Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

“Transposable elements in basidiomycete fungi: dynamics and impact on genome architecture and transcriptional profiles”

Memoria presentada por:

RAÚL CASTANERA ANDRÉS

Para optar al Grado de Doctor por la Universidad Pública de Navarra

Pamplona, 2017

Tesis dirigida por:

Dra. Lucía Ramírez Nasto

Catedrática de Genética y Mejora Vegetal
Grupo de Genética y Microbiología

Dr. LUCÍA RAMÍREZ NASTO, Catedrática de Genética y Mejora Vegetal del Departamento de Producción Agraria de la Universidad Pública de Navarra,

INFORMA:

Que la presente tesis doctoral, "Transposable elements in basidiomycete fungi: dynamics and impact on genome architecture and transcriptional profiles" elaborada por **D. RAÚL CASTANERA ANDRÉS,** ha sido realizada bajo su dirección, y que cumple las condiciones exigidas por la legislación vigente para optar al grado de Doctor.

Y para que así conste, firma la presente en Pamplona, a 8 de Marzo de 2017

Fdo. Lucía Ramírez Nasto

AGRADECIMIENTOS

En primer lugar quiero agradecer a mi directora de tesis Lucía Ramírez, Catedrática de Genética y responsable del grupo GENMIC, primero por darme la oportunidad allá por el año 2011 de formar parte del este grupo de investigación. Más adelante por proponerme la posibilidad de realizar un doctorado, y especialmente por la gran confianza que ha depositado en mí durante todos estos años. Por supuesto, también quiero agradecer a Gerardo Pisabarro, Catedrático de Microbiología. Sin sus consejos difícilmente esta tesis podría haber salido adelante. A los dos tengo que agradecer el haber contado conmigo para contribuir en multitud de proyectos y el haberme introducido en el mundillo de la genómica de hongos, que me ha permitido aprender, viajar y conocer a muchas personas y lugares distintos.

También quiero agradecer a los directores de mis estancias internacionales: Jason Stajich de la Universidad de California Riverside, Igor Grigoriev del Joint Genome Institute y a Hadi Quesneville, del INRA-URGI, por acogerme en sus grupos de investigación.

A mis compañeras y excompañeras Gúmer, Leticia, Marta, Ale, Elaia por todos los momentos dentro y fuera del laboratorio, risas en el café, momentos de “hoy toca comedor” pero “prefiero ir al Sario”, discusiones sobre el estilo de vida navarro, aragonés o argentino, etc.

Como no a Manu, un referente durante mis años por estas tierras, además de un gran compañero de piso y despacho a tiempo parcial.

A Alessandra, con quien he compartido momentos y lugares tan dispares como una excursión entre Sequoias en Muir Woods, una fiesta de Halloween en Los Angeles o un juevincho en Pamplona, además de muchas horas de Skype discutiendo sobre de la metilación de los transposones.

A mis amigos y familia de Zaragoza, Javi, Manu, Ruiz, Fausto, Rubén, Mosteo y muchas otras personas que me dejo pero que saben que son importantes.

A mis abuelos, a mi tía y a mi hermano Diego; he tenido la suerte de poder seguir sus pasos desde que aprendí a andar.

A mis padres, que me han hecho la persona que soy hoy en día

TABLE OF CONTENTS

Resumen

Los elementos transponibles (ETs), también conocidos como elementos móviles o transposones, son unidades genéticas que han desempeñado un importante papel en la evolución de los organismos eucariotas. Su impacto en la arquitectura genómica y en los rasgos fenotípicos observados resultantes de la expresión de dichos genomas se ha estudiado con profundidad en plantas y animales desde su descubrimiento en 1950 por Barbara McClintock. Su capacidad para movilizarse de un *locus* a otro los convierte en herramientas naturales para generar diversidad, ya que conduce a la producción de alteraciones genómicas con efectos deletéreos, neutros o beneficiosos sobre los huéspedes. Por lo tanto, su supervivencia en el genoma depende del equilibrio entre su actividad y la "permisividad" de su huésped. A pesar de que los ETs vegetales y animales han recibido mucha atención, se sabe muy poco sobre su ocurrencia e impacto en el reino de los hongos. De hecho, el primer transposón identificado en un hongo se describió en *Neurospora crassa* en 1989, casi 40 años después del descubrimiento del primer transposón en plantas. Hoy en día, los avances en las tecnologías de secuenciación de ADN han abierto la posibilidad de estudiar el genoma completo de multitud de especies. El número de genomas de hongos secuenciados aumenta a un ritmo sin precedentes, y la mayoría de los esfuerzos se concentran en los basidiomicetos, un grupo de hongos de gran interés debido a su papel en los ecosistemas naturales y a su utilidad en múltiples aplicaciones industriales. En este sentido, la cantidad de información genómica liberada ofrece una oportunidad única para comenzar a descifrar el efecto que los elementos móviles tienen en los genomas de los hongos.

A fecha de inicio de esta tesis de doctorado (Enero de 2013), existía muy poca información sobre transposones en basidiomicetos, ya que la mayoría de las investigaciones se habían concentrado en la caracterización funcional de los genes. A la luz de estos precedentes, el presente trabajo trata sobre la distribución, características e impacto de los transposones en genomas de hongos, con especial énfasis en los basidiomicetos. Se ha utilizado *Pleurotus ostreatus* como modelo de trabajo y se han desarrollado herramientas bioinformáticas para detectar la presencia de transposones a partir de extensos datos genómicos. Este enfoque ha permitido la cuantificación y caracterización del contenido de ETs en numerosos genomas. Además, se ha podido describir el efecto que las inserciones de ETs producen a nivel genómico y transcriptómico.

Esta tesis doctoral se organiza como sigue: En el capítulo I se presenta una introducción que incluye la secuenciación de ADN, el estado actual de la investigación en genómica de hongos y la biología de los elementos transponibles. El capítulo II describe las principales características de los helitrones en basidiomicetos. Los helitrones son un grupo de transposones de ADN que se caracterizan por su mecanismo de transposición de círculo rodante, así como por su capacidad para

capturar y amplificar fragmentos de genes en el genoma del huésped. Sus características y distribución en hongos eran completamente desconocidas al comienzo de este trabajo. El análisis comparativo realizado en dos genomas de *P. ostreatus* mostró la presencia de dos familias de helitrones que interrumpen la colinealidad y originan falta de sintenia. Se identificaron helitrones potencialmente autónomos y con capacidad de transcribirse. Además, algunos elementos contenían genes expresados de origen y funciones desconocidas, así como dominios eucarióticos, bacterianos y virales. La reconstrucción filogenética de los dominios conservados de helitrones eucarióticos reveló su origen polifilético, que podría ser explicado por eventos antiguos de transferencia horizontal. Estos hallazgos fueron el punto de partida para el análisis del contenido de transposones en un amplio rango de especies de hongos, con un enfoque especial en los basidiomicetos y utilizando *P. ostreatus* como modelo.

El capítulo III describe la anotación exhaustiva de ETs realizada en 18 genomas, incluyendo cepas de la misma especie y especies del mismo género. Los resultados indican un escenario de excepcional variabilidad, ya que se encontraron especies cuyo genoma estaba ocupado entre el 0,02 y el 29,8% por elementos transponibles. Un análisis detallado realizado sobre dos cepas de *Pleurotus ostreatus* descubrió un genoma ocupado principalmente por elementos de ARN, especialmente retrotransposones que han mostrado ser activos durante los últimos dos millones de años. La acumulación preferencial de estos retrotransposones ha conducido a la aparición de regiones genómicas que carecen de conservación tanto a nivel intra como inter-específico. Además, se estudió el efecto de las inserciones de transposones en la expresión de los genes cercanos. Los resultados demuestran que la expresión de un número importante de dichos genes está silenciada, observándose una represión más fuerte cuando los genes se localizaron dentro de las regiones enriquecidas en transposones. El análisis transcripcional realizado en cuatro especies de hongos reveló que este silenciamiento mediado por ETs estaba presente sólo en especies con maquinaria activa de metilación de citosinas, lo que sugiere que este fenómeno podría estar relacionado con mecanismos de defensa epigenética dirigidos a evitar la proliferación de los elementos móviles. Todos los análisis descritos anteriormente fueron posibles debido a la disponibilidad pública de secuencias genómicas y anotaciones realizadas por consorcios internacionales. En este sentido, el grupo de investigación de Genética y Microbiología (GENMIC) ha contribuido liderando la secuenciación y anotación de *Coniophora olivacea*, un basidiomiceto de podredumbre parda del orden Boletales. En el capítulo IV describe el resultado de la secuenciación y anotación del genoma, así como el análisis comparativo con otras especies de su mismo orden. Dichos análisis revelaron la presencia de expansiones genómicas diferenciales en el orden Boletales, causadas por ráfagas de amplificación de retrotransposones en el curso de la evolución. Finalmente, en el capítulo V se discuten los resultados de esta tesis doctoral en el contexto de la literatura más reciente,

contribuyendo a una mejor comprensión del impacto de ETs en los basidiomicetos de acuerdo a su filogenia y estilo de vida. Se discute la fuerte influencia de los ETs sobre el tamaño del genoma, así como su papel diferencial en la evolución de los patógenos, simbiontes y hongos ligninolíticos. Por último, se proporcionan ejemplos de las formas en que la información publicada en esta tesis puede aplicarse a la investigación en hongos y a la biotecnología industrial.

Summary

Transposable elements (TE), also known as mobile elements or transposons, are enigmatic genetic units that have played important roles in the evolution of eukaryotes. Their impact on genome architecture and phenotypic traits has been widely studied in plants and animals, ever since their discovery in maize during the 1950s by Barbara McClintock. Their ability to move from one locus to another makes them natural tools for generating diversity, as this characteristic leads to genomic alterations with deleterious, neutral, or beneficial effects on hosts. Thus, their survival in the genome depends on the equilibrium between their own benefit and their host's "permissibility." Plant and animal TEs have received much attention, yet very little is known about their occurrence and impact on the fungal kingdom. In fact, the first fungal TE was described in *Neurospora crassa* in 1989, about 40 years later than the discovery of the first TE in plants. Today, revolutionary advances in genome sequencing have opened the possibility of studying non-model species at a whole-genome level. The number of fungal-sequenced genomes increases daily at an unprecedented rate, and most efforts are being concentrated on basidiomycetes, a group of fungi of great interest due to their role in natural ecosystems and their use in multiple industrial applications. In this sense, the amount of genomic information released offers a unique opportunity to start deciphering the effect that mobile, repetitive elements have on fungal genomes.

At the time of the start of this PhD thesis (January 2013), very little information regarding basidiomycete TEs existed, as most research was focused on the functional characterization of protein-coding genes. In light of these precedents, the main topics covered in this work are the distribution, characteristics, and impact of transposons in fungal genomes, with an emphasis on basidiomycetes. Using *Pleurotus ostreatus* as a working model, bioinformatics pipelines have been developed to dig into the extensive genomic data to obtain high quality TE annotations. This approach has allowed for the quantification and characterization of the transposon load of many fungal species and for the testing of hypotheses about the effect that TE insertions produce at the genomic and transcriptomic level.

This PhD thesis is organized as follows. **Chapter I** introduces genome sequencing as well as current state-of art research on fungal genomics and transposable elements biology. **Chapter II** describes the main characteristics of basidiomycete helitrons. Helitrons are a unique group of DNA transposons, characterized by a proposed rolling-circle transposition mechanism as well as an ability to capture and amplify gene fragments across the host genome. Their characteristics and distribution in fungi were completely unknown at the onset of this work. Comparative analysis performed in two *P. ostreatus* genomes showed the presence of two helitron families that disrupt

gene co-linearity and cause important lack of synteny. Putative autonomous helitrons that were transcriptionally active were identified. Some carried highly expressed captured genes of unknown origin and function. In addition, both helitron families contained eukaryotic, bacterial, and viral domains within their boundaries. A phylogenetic reconstruction of the conserved domains of eukaryotic helitron-encoded helicases revealed a polyphyletic origin, which might be explained by ancient horizontal transfers. These findings were the starting point for the analysis of the whole TEs landscape in a wide range of species across the fungal phylogeny, with a special focus on basidiomycetes and using *P. ostreatus* as a model. In this regard, **Chapter III** describes an exhaustive TE annotation performed in 18 genomes, including strains of the same species and species of the same genera. Our results depicted a scenario of exceptional variability, wherein species have from 0.02 to 29.8% of their genome consisting of transposable elements. A detailed analysis performed on two strains of *Pleurotus ostreatus* uncovered a genome populated mainly by class I elements, especially LTR-retrotransposons result of recent bursts of amplification (0 to 2 million years ago). The preferential accumulation of TEs into clusters led to the presence of genomic regions that lack intra- and inter-specific conservation. In addition, the effect of TE insertions on the expression of their nearby upstream and downstream genes was also studied. Results showed that an important number of genes that were under TE influence were significantly repressed. In addition, stronger repression was observed when genes were localized within transposon clusters. The transcriptional analysis performed in four additional fungal species revealed that this TE-mediated silencing was present only in species with active cytosine methylation machinery, suggesting that this phenomenon might be related to epigenetic defense mechanisms aimed at controlling TE proliferation. All of the analyses previously described were possible due to the public availability of genome sequences and gene annotations carried out by international consortia. With the aim of contributing to this effort, the Genetics and Microbiology research group (GENMIC) led the sequencing and annotation of *Coniophora olivacea*, a brown rot basidiomycete of the Boletales order. **Chapter IV** describes the comparative analysis of *C. olivacea* with other Boletales, which revealed the presence of species-specific genome expansions caused by TE amplification bursts at different time points in the course of evolution. Finally, **Chapter V** discusses this PhD thesis' findings in the context of the most recent related literature, contributing to a better understanding of the impact of TEs in basidiomycetes according to their phylogeny and lifestyle. The strong influence of TEs on basidiomycetes' genome size as well as their differential role in the evolution plant pathogens, symbionts, and wood decayers is discussed. Finally, examples of the ways in which the information released in this thesis can be applied to fungal research and industrial biotechnology are provided.

Objectives

The main objectives of this thesis are:

- Identification, functional characterization, and transcriptional analysis of helitron transposons in *Pleurotus ostreatus.*
- Development of bioinformatics strategies for annotating and analyzing transposable elements in fungal genome assemblies.
- Analysis of the distribution and impact of transposable elements in genome architecture and transcriptional profiles in fungi, with an emphasis on basidiomycetes.
- Sequencing, assembly, annotation, and comparative analysis of the brown-rot basidiomycete *Coniophora olivacea* with related Boletales.

Chapter I: General Introduction

1.1. A tale of DNA sequencing

The possibility of determining the nucleotide composition of genes and genomes revolutionized biological research and fueled the emergence of genomics science. The first gene sequence (that of a yeast alanine transfer RNA) was published in 1965 (Holley et al.). Soon after, the first sequence of a protein-coding gene was described (Jou et al. 1972), and finally the complete sequence of the bacteriophage MS2 RNA was characterized in 1976 (Fiers et al.). The development of techniques such as Maxam–Gilbert's (Maxam and Gilbert 1977) and especially Sanger's (Sanger and Coulson 1975; Sanger et al. 1977) opened the possibility of sequencing longer fragments of DNA more efficiently. Later, the deployment of the Whole Genome Shotgun technique (WGS) led to a faster and more economical method of genome sequencing the genome. These approaches are today referred to as first-generation DNA sequencing technologies.

The workflow of WGS starts by randomly fragmenting the DNA into smaller pieces, followed by cloning and sequencing these fragments, and then further assembly into longer pieces by searching for overlaps. This approach required the development of specific computer softwares, which were used to sequence some of the first complete viral and bacterial genomes (Staden 1979).

Second-generation sequencing appeared in 1988, and was a totally new technology of sequencing DNA that was based on the measurement of the pyrophosphate generated by the DNA polymerization reaction (Hyman 1988). This technique offered the advantage of obtaining the information of the nucleotides without the need to use electrophoresis, and evolved into 454 technology, which produced the first high throughput sequencing machines in the market. Other technologies such Solexa/Illumina, SOLiD and Ion Torrent grew rapidly, competing for the market hegemony. Up to day, the high-throughput and low cost of Illumina sequencing has made it the most successful technology, and it is currently the most habitual choice for generating draft genomes, re-sequencing, and for other applications such as RNAseq or metagenomics. Nevertheless, the small size of the reads produced (i.e., 36-250bp) is still a limitation for assembling repetitive regions of complex genomes.

Third-generation sequencing technologies have overcome this problem, producing rutinarily reads of 3 to 15 kb (Lee et al. 2016), which facilitates the production of highly accurate *de novo* genomes. As a result, technology has changed the way we study biology and evolution, from the analysis of single genes to the study of species from a whole genome perspective.

1.2. From genes to genomes: Toward an understanding of genome complexity

The small genome of the bacteria *Haemophilus influenzae* was the first free-living organism to be sequenced and annotated (Fleischmann et al. 1995). Later, model organisms from different lineages of the eukaryotic domain, such as *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Caenorhabditis elegans* (Equence et al. 1998), and *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) were the prelude to one of the biggest scientific achievements ever carried out: determining the complete 3,000 Mb of sequence of the human genome (Consortium 2001; Venter et al. 2001). Completing such a project was possible due to enormous economical and collaborative efforts from research groups all over the world.

Today, with the evolution of the sequencing technologies, genome projects are no longer the exclusive domain of large collaborative consortia. In fact, the significant reduction in sequencing costs have made possible to produce genome drafts for multiple non-model species. The assembled scaffolds of a genome draft have linear correspondence with fractions of the species chromosomes. If a genome assembly is complete, then each scaffold corresponds to an entire chromosome, providing the researchers with a valuable physical map of the genome of species of interest. In order to extract meaningful information from the assembly, it is necessary to carry out a genome annotation, which involves the use of complex pipelines that combine multiple gene predictions based on *de novo*, homology, and transcriptome information (Thibaud-Nissen et al. 2013; Kuo et al. 2014). The road-map of a genome project has evolved during the last decade, and sequencing a species genome is now an affordable task for a single research group (Fig 1). This evolution has been accompanied by strong progress in the development of bioinformatics tools and in the availability of computing resources, both of which are essential to carrying out downstream analyses such as assembly, similarity searches, or data integration.

Once the first whole genome sequences and annotations became available, scientists quickly realized that genome size and gene content were not as tightly related as initially thought. In fact, the human genome was ~ 250-fold bigger than that of the budding yeast *S. cerevisiae* (3,000 Mb vs 12 Mb), whereas its gene content was just ~ 4-fold (24,000 *vs* 6,200) in size, despite their different complexity. Comparative studies of eukaryotic and prokaryotic genomes suggested that the percentage of coding DNA diminished progressively with increasing genome size, in contrast to non-coding DNA, which was the responsible for the differences found in genome size between cellular species (Lynch and Conery 2003).

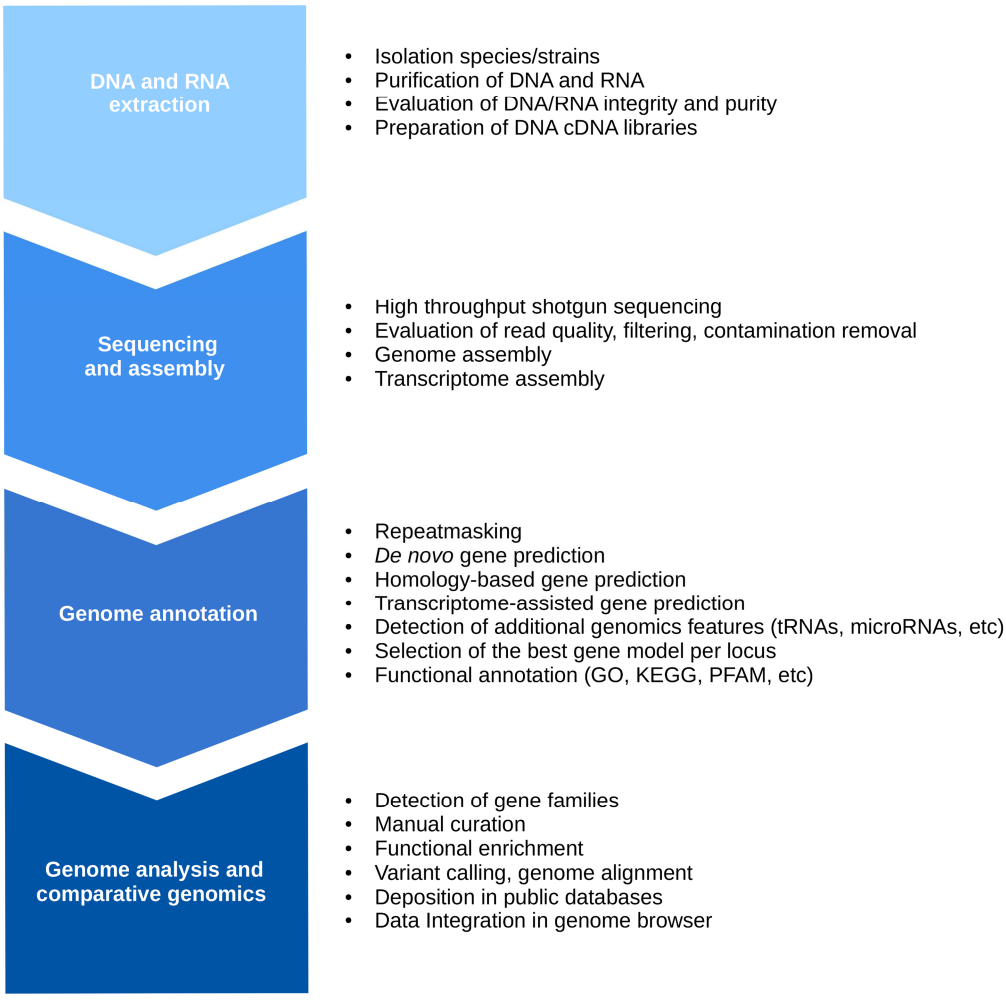

Figure 1. Simplified workflow of a typical genome sequencing project.

1.3. The fungi

The fungal kingdom comprises an ancient group of organisms that diverged from animals about 1500 Million years ago (Mya) (Hedges et al. 2004). Fungi are unique among eukaryotes in that they are able to externally digest food and incorporate nutrients through their cell wall. They reproduce by sexual or asexual spores, and can grow as unicellular organisms or form a multicellular mycelium composed of branching tubular cells called hyphae. Also, they display multiple lifestyles and nutritional modes such as biotrophy, saprotrophy, or necrotrophy. Their growth is associated with plants and animals or they live as free-living organisms in the soil. They are involved in a myriad of processes in natural ecosystems, being key players in nutrient cycling as

primary decomposers. In addition to their importance for natural ecosystems, they have a strong impact on cultivated crops, as well as in animal and human health. Also, fungi are of great interest for biotechnology (ie, for enzyme production), pharmaceuticals (ie, for the production of antibiotics, heterologous expression of vaccines), and the food industries (ie, for the production of mushrooms, cheese or wine).

Fungi have been traditionally classified into four *phyla*: Chytridiomycota, Zygomycota, Ascomycota, and Basidiomycota (Alexopoulos et al. 1996). Nevertheless, a recent consortium of taxonomists proposed the re-classification of the first two (Chytridiomycota and Zygomycota, often referred as early diverging fungi) in up to six new phyla and four unplaced subphyla (Hibbett et al. 2007). Also, they proposed classifying the clade formed by Ascomycota and Basidiomycota as the sub-kingdom Dikarya, which comprises up to 98% of the described fungal species (James et al. 2006). Unlike most eukaryotes, for species belonging to Dikarya the plasmogamy and karyogamy phases of the life cycle occur in distinct phases, separated by the dikaryotic stage. During this stage, the two genomic copies are kept separate in two parental nuclei who share the same cytoplasm. This condition terminates when karyogamy occurs, immediately prior to the onset of the meiotic divisions that produces the haploid monokaryotic spores (Fig 2).

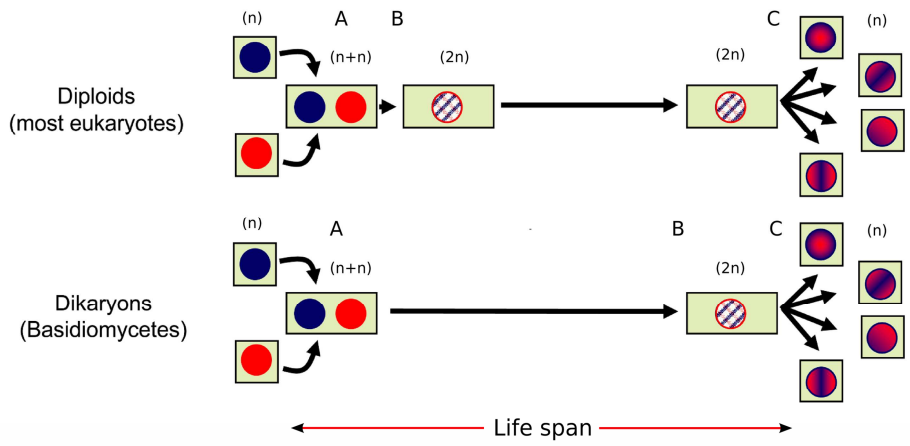

Figure 2. Diploid *versus* dikaryotic life cycle (adapted from Castanera et al. 2013). Cells are represented as rectangles with circles (nuclei) inside. Plasmogamy is marked by (A), karyogamy by (B), and meiosis by (C). In diploids, A and B occur simultaneously or with a slight delay. In dikaryons, A and B are separated by a different period of time.

By contrast, in diploids karyogamy occurs during or shortly after plasmogamy, and the two equivalent genomic copies present in the cell are fused into a single nucleus. The total genetic complement of dikaryons (n+n) and diploids (2n) is the same, yet the special organization of dikaryons creates different behavior and evolutionary expectations (Anderson and Kohn 2007). The main conserved difference between members of the two *phyla* of Dikarya is the cell in which karyogamy and meiosis occur: Ascomycota species produce sexual spores in sac-shaped organs known as asci, whereas Basidiomycota species produce them in the basidia. Ascomycota is the largest fungal phyla and contains some of the most relevant fungal models, such as *Saccharomyces cerevisiae* or *Neurospora crassa.* Basidiomycota contains the mushroom-forming species, some of which are of great importance for the food industry, such as *Agaricus bisporus* or *Pleurotus ostreatus.*

1.4. Genomics of basidiomycete fungi

Basidiomycete fungi represent an extraordinarily diverse group of organisms that have different lifestyles. This phylum is constituted by three sub-phylla: *Agaricomycotina*, *Ustilaginomycotina*, and *Pucciniomycotina* (Fig 3).

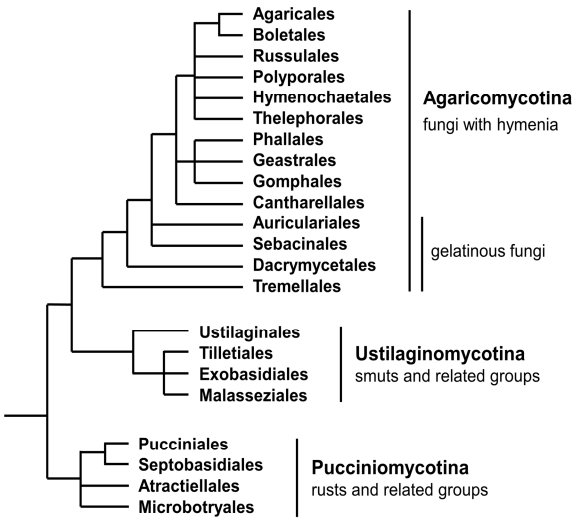

Figure 3. Phylogeny of basidiomycetes. Obtained from Piepenbring (2015).

Beyond the popularity of edible mushrooms belonging to *Agaricomycotina*, ligninolytic species have been most studied during the last decade, due to for their potential industrial and biotechnological applications (Floudas et al. 2012; Alfaro et al. 2014). Additionally, plant pathogens such smuts (*Ustilaginomycotina*) and rusts (*Pucciniomycotina*) have received great interest because of their impact on agricultural crops. Basidiomycetes also include saprophytes, mycorrhizal species, animal pathogens, and endophytes, and they show morphological diversity, with species displaying filamentous and yeast forms. The ways in which this huge diversity is reflected in specific genomic attributes is a key question that challenges the current understanding of fungal genomics and biology. With the advent of next generation sequencing (NGS) techniques, hundreds of fungal genomes have been sequenced and annotated. The development of public databases such as Mycocosm (Grigoriev et al. 2014) or FungiDB (Stajich et al. 2012), which include genome browsers and downloadable data presented in a consistent and organized fashion, has fueled the study of the comparative genomics of fungi. Since the release of the first basidiomycete genome sequence in 2004 (*Phanerochaete chrysosporium*, Martinez et al. 2004), international efforts have contributed to the production of more than 100 additional basidiomycete genome annotations (Floudas et al, 2012; Kohler et al, 2015) (http://genome.jgi.doe.gov/basidiomycota/ basidiomycota.info.html). Genomic data obtained from these sequencing projects have shed some light on the main features of fungal (and particularly basidiomycetes) genomes, which are small in comparison to those of plants and animals, with small inter-genic distances and short introns. Specifically, the variability found in this first set of genomes is intriguing, reaching up to 21-fold genome size variations (ie, 8.9 Mb for *Malassezia globosa vs* 189.5 Mb for *Melampsora lini*) (Xu et al. 2007; Nemri et al. 2014). From studies of other eukaryotic models, we know that genome size variability may be related to expanded non-coding DNA such introns and repetitive DNA, and especially due to the amplification of TEs.

1.5. Genetics and genomics of *Pleurotus ostreatus*

Pleurotus ostreatus is an edible basidiomycete that is of great importance for the food industry due to its excellent organoleptic properties. In its natural environment it grows on tree stumps, causing white-rot decay, although it can be easily cultivated on artificial substrates under controlled conditions. In the last two decades, *P. ostreatus* has gained a special relevance in the biotechnology industry, as the wide set of non-specific enzymes that it uses for wood degradation are also able to oxidize a broad spectrum of recalcitrant compounds related to lignin (Novotný et al. 2004). This makes it a promising agent for bioremediation, pulp bleaching, and second-generation bioethanol

production, among other applications (Kunamneni et al. 2008).

During the last two decades, the Genetics and Microbiology research group from the Public University of Navarre (GENMIC) has used this species as a model to study basidiomycete genetics and genomics. The first studies on *P. ostreatus* genetics described the molecular karyotype of the commercial strain N001, its genetic linkage map, and a number of Quantitative Trait Loci (QTL) linked to agronomic and industrially-relevant regions such as mushroom yield, earliness, or enzyme production (Larraya et al. 1999; Larraya et al. 2003; Santoyo et al. 2008). With the advent of NGS technologies, all these efforts crystallized into an international genome sequencing project that led to the sequencing, assembly, and annotation of the two N001 parental monokaryotic strains PC15 and PC9 (Fig 4).

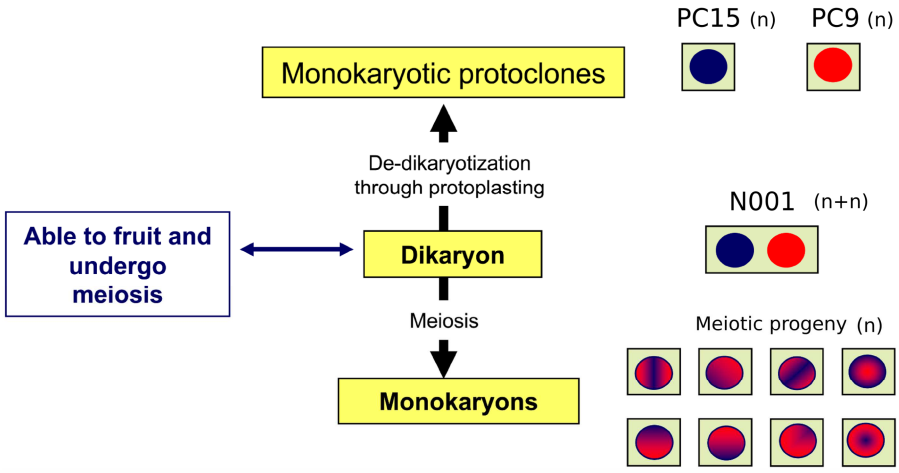

Figure 4. Schematic representation of the origin of *P. ostreatus* strains used in this thesis (adapted from Castanera et al., 2013).

PC15 was sequenced with the Sanger whole-genome shotgun approach, and PC9 was sequenced using the Sanger whole genome shotgun and 454 paired-end sequencing reads (Riley et al. 2014). PC15 genome assembly version 2.0 (34.3 Mb) was subjected to targeted genome improvement, which led to the complete assembly of 12 scaffolds with a very low gap content (a single 96 base-pairs gap in the whole assembly) that matched the corresponding *P. ostreatus* chromosomes (11 nuclear plus 1 mitochondrial chromosome). In contrast, PC9 assembly v1.0 (35.6 Mb) contains 572 scaffolds and a total of 476 gaps that cover 9.72 % of the whole assembly. Both genome assemblies and annotations are publicly available in the Mycocosm database (http://genome.jgi.doe.gov/). The preliminary analyses showed a very high conservation in gene content. By contrast, the synteny

between both genomes is frequently interrupted by inversions, translocations, or deletions, similar to those promoted by transposable elements in other eukaryotes.

1.6. Biology of transposable elements

Repetitive DNA sequences are essential components of eukaryotic genomes. Depending on their nature and characteristics they are classified in simple sequence repeats (SSR, which include mini- and microsatellites), transposable elements, multi-copy genes (including tRNAs, sRNA and rRNA) and integrated viruses (Jurka 2000). Within these categories, transposable elements are the most abundant and complex elements of eukaryotic genomes. TEs are mobile genetic units that colonize genomes and generate intra- and inter-specific variability, producing profound structural (i.e., genome rearrangements and gene mutations) and functional (i.e., affect gene expression) genome alterations. Despite the ubiquity of TEs in the eukaryotic domain, the genome fraction occupied by these elements is highly diverse, accounting for approximately 3 % in yeasts (Kim et al. 1998), up to 50 % in mammalian genomes (Zamudio and Bourc'his 2010) and more than 80 % in some plants, including wheat or maize (Schnable et al. 2009; Wicker et al. 2011). The expansion of these elements is mediated by transposition events that can lead to their own duplication. TEs are classified into two classes based on their transposition mechanisms. Class I elements transpose via RNA intermediates and include five orders (LTR, DIRS, PLE, LINE, and SINE) that are differentiated based on their structure and transposition system. Class II encompasses elements that transpose directly from DNA to DNA. This class is divided into two subclasses: One includes the TIR and Crypton orders, and the other contains Helitrons and Mavericks. In addition, TE families are formed by both autonomous (coding for the proteins necessary for its transposition) and non-autonomous elements that rely on compatible transposases/retrotransposases for their mobilization. The majority of transposable elements generate target site duplications at their insertion sites (TSD), which are formed as part of the insertion process. Exceptions include Helitrons (Kapitonov and Jurka 2001) and the recently discovered Spy elements (Han et al. 2014).

1.7. Transposable elements and eukaryotic genome defense

From the perspective of a mobile element family, its success depends critically on the ability to increase in copy number without risking host viability, in order to increase the possibility to be transmitted to the progeny by vertical inheritance. From the perspective of the host genome, TEs are

a source of instability due to their mutagenic effect. Nevertheless, its presence and activity in certain regions is essential for genome integrity. Examples to this are telomeres and centromeres, which are built by microsatellites, nested blocks of TEs and other repeats (Kumekawa et al. 2001; Gao et al. 2015). Besides natural selection may act against the maintenance of aggressive TEs, studies in plants, animals and fungi have shown that all of them have developed host-encoded epigenetic mechanisms to control their proliferation. These mechanisms are known as TGS (Transcriptional Gene Silencing) and PTGS (Post-Transcriptional Gene Silencing). In animal genomes, TGS operates through the production of Piwi-interacting RNAs (piRNAs) that lead to heterochromatin formation and transcriptional silencing of the TEs. Intriguingly, this mechanism is absent in plants and fungi. Chromatin modifications have been also described to be a very efficient way to shut down TE expression. Especially DNA methylation, which has been described to silence TEs in plants (Saze et al. 2012) and animals (Chen et al. 1998). PTGS targets TEs in an homology-dependent manner. By this mechanism, aberrant double-stranded RNAs (dsRNA) produced by TEs are cleaved by Dicer protein, producing small interfering RNAs (siRNAs) that guide RNA-degrading complexes to a complementary transcript. Both TGS and PTGS are interconnected, as they are related directly or indirectly to the RNAi machinery. In fact, in plants and mammals DNA-methylation appears to be guided by small RNAs produced by the RNAi pathway (Zilberman et al. 2003; Kawasaki and Taira 2004). In fungi, in addition to the epigenetic inactivation of TEs by DNA methylation (Montanini et al. 2014) and RNAi (Dang et al. 2011), a third homology-dependent mechanism has been described, called RIP (Repeat-induced point mutation). This process produces an hypermutation of repeated DNA by promoting G:C to A:T transitions in the sexual phase (Selker et al. 1987). The complex epigenetic control of TEs in eukaryotes promotes a side-effect that has important consequences for genome functioning: TEs shift from being genomic parasites to be genome regulators. This field has been studied in plants and mammals, but is totally obscure in fungi. In this sense, an increasing number of examples describe how TEs can modify the expression of surrounding genes by the generation of epialleles in plants (Iida et al. 2004) or mammals (Morgan et al. 1999), and even regulate phenotypic traits in response to stress and environmental factors (Capy et al. 2000).

1.8. Transposable elements in basidiomycete fungi

In the 1990s, several studies identified transposon-like sequences in some basidiomycete species. These studies highlighted their potential impact on genomic variability and gene regulation (Gaskell et al. 1995; Sonnenberg et al. 1999). Soon after, *Scooter*, the first active DNA transposon was

described in a member of the phylum *Basidiomycota* (Fowler and Mitton 2000). It was also found that this phylum contained multiple species ferrying the LTR (Long-Terminal Repeat) retrotransposon *marY1* (Murata H et al. 2001). The importance of class I TEs in Basidiomycetes was emphasized when the genome sequence of *Phanerochaete chrisosporium*, a lignin degrading fungus, was published in 2004 (Martinez et al. 2004). This study revealed the first hints of the genome-wide TE landscape and reported several insertions of class I TEs in genes involved in lignin degradation. Today, approximately 100 basidiomycete genomes have been sequenced and published, and with TE annotation procedures becoming more mature, the opportunities to study these repetitive, enigmatic sequences have increased exponentially. Nevertheless, the study of repeat sequences in most genome projects is usually minimized in comparison to other kind of analysis, and very often valuable information about their presence, structure or impact in the genome is kept in the dark.

1.9. References

Alexopoulos CJ, Mims CW, Blackwell M (1996) Introductory Mycology. 4th Edition. Wiley, New York 868. doi: 10.2105/AJPH.43.6_Pt_1.781-a

Alfaro M, Oguiza JA, Ramírez L, Pisabarro AG (2014) Comparative analysis of secretomes in basidiomycete fungi. J Proteomics 102:28–43. doi: 10.1016/j.jprot.2014.03.001

Anderson JB, Kohn LM (2007) Dikaryons, diploids, and evolution. Sex in Fungi. ASM Press Washington, DC 333–348. doi: 10.1128/9781555815837.ch20

Castanera R, Omarini A, Santoyo F, Pérez G, Pisabarro AGAG, Ramírez L (2013) Non-additive transcriptional profiles underlie dikaryotic superiority in *Pleurotus ostreatus* laccase activity. PLoS One 8:e73282. doi: 10.1371/journal.pone.0073282

Capy P, Gasperi G, Biémont C, Bazin C (2000) Stress and transposable elements: co-evolution or useful parasites? Heredity (Edinb) 85:101–106. doi: 10.1046/j.1365-2540.2000.00751.x

Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R (1998) DNA hypomethylation leads to elevated mutation rates. Nature 395:89–93. doi: 10.1038/25779

Consortium IHGS (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. doi: 10.1038/35057062

Dang Y, Yang Q, Xue Z, Liu Y (2011) RNA interference in fungi: Pathways, functions, and applications. Eukaryot. Cell 10:1148–1155. doi: 10.1128/EC.05109-11

Equence CES, Iology TOB, The C, Consortium S, Consortium TC elegans S (1998) Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. Science 282:2012–2018. doi: 10.1126/science.282.5396.2012

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers a, Van den Berghe a, Volckaert G, Ysebaert M (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260:500–507. doi: 10.1038/260500a0

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512. doi: 10.1126/science.7542800

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, De Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, John FS, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev I V, Hibbett DS (2012) The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science (80) 336:1715–1719. doi: 10.1126/science.1221748

Fowler TJ, Mitton MF (2000) Scooter, a new active transposon in *Schizophyllum commune*, has disrupted two genes regulating signal transduction. Genetics 156:1585–1594.

Gaskell J, Van den Wymelenberg A, Cullen D (1995) Structure, inheritance, and transcriptional effects of Pce1, an insertional element within *Phanerochaete chrysosporium* lignin peroxidase gene lipI. Proc Natl Acad Sci U S A 92:7465–7469. doi: 10.1073/pnas.92.16.7465

Gao D, Jiang N, Wing RA, Jiang J, Jackson SA (2015) Transposons play an important role in the evolution and diversification of centromeres among closely related species. Front Plant Sci 6:216. doi: 10.3389/fpls.2015.00216

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 Genes. Science (80) 274:546–567. doi: 10.1126/science.274.5287.546

Grigoriev I V, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res 42:D699–D704. doi: 10.1093/nar/gkt1183

Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2. doi: 10.1186/1471-2148-4-2

Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., Lumbsch, T., Lutzoni, F., Matheny, P.B., McLaughlin, D.J., Powell, M.J., Redhead, S., Schoch, C.L., Spatafora, J.W., N (2007) A higher-level phylogenetic classification of the Fungi. Mycol Res 111:509–547. doi: 10.1016/j.mycres.2007.03.004

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A (1965) Structure of a ribonucleic acid. Science 147:1462–5. doi: 10.1126/science.147.3664.1462

Hyman ED (1988) A new method of sequencing DNA. Anal Biochem 174:423–36. doi:10.1073/pnas.74.2.560

Iida S, Morita Y, Choi JD, Park KI, Hoshino A (2004) Genetics and epigenetics in flower pigmentation associated with transposable elements in morning glories. Adv Biophys 38:141–159. doi: 10.1016/S0065-227X(04)80136-9

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung G-H, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüßler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443:818–822. doi: 10.1038/nature05110

Jou WM, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. Nature 237:82–88. doi: 10.1038/237082a0

Jurka J (2000) Repbase Update - a database and an electronic journal of repetitive elements. Trends Genet 16:418–420. doi: 10.1016/S0168-9525(00)02093-X

Kapitonov V V, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98:8714–8719. doi: 10.1073/pnas.151269298

Kawasaki H, Taira K (2004) Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. Nature 431:211–217. doi: 10.1038/nature02783

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8:464–478. doi: 10.1101/gr.8.5.464

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canbäck B, Choi C, Cichocki N, Clum

A, Colpaert J, Copeland A, Costa MD, Doré J, Floudas D, Gay G, Girlanda M, Henrissat B, Herrmann S, Hess J, Högberg N, Johansson T, Khouja H-R, LaButti K, Lahrmann U, Levasseur A, Lindquist EA, Lipzen A, Marmeisse R, Martino E, Murat C, Ngan CY, Nehls U, Plett JM, Pringle A, Ohm RA, Perotto S, Peter M, Riley R, Rineau F, Ruytinx J, Salamov A, Shah F, Sun H, Tarkka M, Tritt A, Veneault-Fourrey C, Zuccaro A, Tunlid A, Grigoriev I V, Hibbett DS, Martin F (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet 47:410–5. doi: 10.1038/ng.3223

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H (2001) The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 4. DNA Res 8:285–90.

Kunamneni A, Plou FJ, Ballesteros A, Alcalde M (2008) Laccases and their applications: a patent review. Recent Pat Biotechnol 2:10–24. doi: 10.2174/187220808783330965

Kuo A, Bushnell B, Grigoriev I V (2014) Fungal genomics: Sequencing and annotation. Adv. Bot. Res. 70:1–52.

Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S, McCombie WR, Schatz M (2016) Third-generation sequencing and the future of genomics. Biorxiv doi:https://doi.org/10.1101/048603

Lynch M, Conery JS (2003) The origins of genome complexity. Science (80-) 302:1401–1404. doi: 10.1126/science.1089370

Martinez D, Larrondo LF, Putnam N, Gelpke MDS, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, Rokhsar D (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat Biotechnol 22:695–700. doi: 10.1038/nbt967

Maxam M, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74:560–4. doi: 10.1073/pnas.74.2.560

Montanini B, Chen P-YY, Morselli M, Jaroszewicz A, Lopez D, Martin F, Ottonello S, Pellegrini M (2014) Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. Genome Biol 15:411. doi: 10.1186/s13059-014-0411-5

Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23:314–318. doi: 10.1038/15490

Murata H, Miyazaki Y, Babasaki K (2001) The Long Terminal Repeat (LTR) sequence of marY1, a retroelement from the ectomycorrhizal homobasidiomycete *Tricholoma matsutake*, is highly conserved in various higher fungi. Bioscience, Biotechnology, and Biochemistry 65(10):2297-2300. doi: 10.1271/bbb.65.2297

Nemri A, Saunders DG, Anderson C, Upadhyaya NM, Win J, Lawrence GJ, Jones DA, Kamoun S, Ellis JG, Dodds PN (2014) The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. Front Plant Sci 5:98. doi: 10.3389/fpls.2014.00098

Novotný Č, Svobodová K, Erbanová P, Cajthaml T, Kasinath A, Lang E, Šašek V (2004) Ligninolytic fungi in bioremediation: Extracellular enzyme production and degradation rate. In: Soil Biology and Biochemistry. pp 1545–1551

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–448. doi: 10.1016/0022-2836(75)90213-2

Sanger F, Nicklen S, Coulson a R (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–7. doi: 10.1073/pnas.74.12.5463

Saze H, Tsugane K, Kanno T, Nishimura T (2012) DNA methylation in plants: Relationship to

small rnas and histone modifications, and functions in transposon inactivation. Plant Cell Physiol. 53:766–784. doi:10.1093/pcp/pcs008

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science (80-) 326:1112–1115. doi: 10.1126/science.1178534

Selker EU, Cambareri EB, Jensen BC, Haack KR (1987) Rearrangement of duplicated DNA in specialized cells of *Neurospora*. Cell 51:741–752. doi: 10.1016/0092-8674(87)90097-3

Sonnenberg ASM, Baars JJP, Mikosch TSP, Schaap PJ, Van Griensven LJLD (1999) Abr1, a transposon-like element in the genome of the cultivated mushroom *Agaricus bisporus* (Lange) imbach. Appl Environ Microbiol 65:3347–3353.

Staden R (1979) A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 6:2601–2610. doi: 10.1093/nar/6.7.2601

Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, Harb OS, Kissinger JC, Li W, Nayak V, Pinney DF, Stoeckert Jr. CJ, Roos DS (2012) FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res 40:D675-81.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815. doi: 10.1038/35048692

Thibaud-Nissen F, Souvorov A, Murphy T, DiCuccio M, Kitts P (2013) Eukaryotic Genome Annotation Pipeline. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK169439/

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov G V, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang

X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander K V, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X, Sinsheimer RL, Sanger F, Seeburg PH, Strauss EC, Kobori JA, Siu G, Hood LE, Gocayne J, Martin-Gallardo A, McCombie WR, Jensen MA, Adams MD, Adams MD, Adams MD, Kerlavage AR, Fields C, Venter JC, Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC, Polymeropoulos MH, Marra M, Adams MD, White O, Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB, Mahy BWJ, Esposito JJ, Venter JC, Fleischmann RD, Fraser CM, Bult CJ, Tomb JF, Klenk HP, Venter JC, Smith HO, Hood L, Schmitt H, Zhao S, Lin X, Weber JL, Myers EW, Green P, Pennisi E, Venter JC, Adams MD, Marshall E, Pennisi E, Adams MD, Rubin GM, Myers EW, Collins FS, Sanger F, Nicklen S, Coulson AR, Prober JM, Myers G, Selznick S, Zhang Z, Miller W, Hattori M, Dunham I, Carvalho AB, Lazzaro BP, Clark AG, Schuler GD, Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Olivier M, Chaudhari N, Hahn WE, Milner RJ, Sutcliffe JG, Dickson D, Ewing B, Green P, Crollius HR, Pruitt KD, Katz KS, Sicotte H, Maglott DR, Uberbacher EC, Xu Y, Mural RJ, Burge C, Karlin S, Mural RJ, Salamov AA, Solovyev V V., Miklos GL, John B, Francke U, Horvath JE, Schwartz S, Eichler EE, Bickmore WA, Sumner AT, Holmquist GP, Bernardi G, Zoubak S, Clay O, Bernardi G, Ohno S, Broman KW, Murray JC, Sheffield VC, White RL, Weber JL, McEachern MJ, Krauskopf A, Blackburn EH, Bird A, Gardiner-Garden M, Frommer M, Larsen F, Gundersen G, Lopez R, Prydz H, Cross SH, Bird A, Grunau C, Hindermann W, Rosenthal A, Antequera F, Bird A, Cross SH, Slavov D, Smit AF, Riggs AD, Elliott DJ, Makeyev A V., Chkheidze AN, Lievhaber SA, Pan Y, Decker WK, Huq AHHM, Craigen WJ, Nouvel P, Goncalves I, Duret L, Mouchiroud D, Smith TF, Waterman MS, Delcher AL, Trask BJ, Sharon D, Barbazuk WB, McLysaght A, Enright AJ, Skrabanek L, Wolfe KH, Burt DW, Skrabanek L, Wolfe KH, Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY, Taillon-Miller P, Piernot EE, Kwok PY, Altshuler D, Marth GT, Cargill M, Halushka MK, Zhang J, Madden TL, Nachman MW, Bauer VL, Crowell SL, Aquadro CF, Nickerson DA, Jorde L, Wang DG, Przeworski M, Hudson RR, Rienzo A Di, Tavare S, Clark AG, Kaessmann H, Heissig F, Haeseler A von, Paabo S, Sonnhammer EL, Eddy SR, Durbin R, Bateman A, Ponting CP, Schultz J, Milpetz F, Bork P, Goodenough DA, Goliger JA, Paul DL, Wilkinson DG, Nakamura F, Kalb RG, Strittmatter SM, Horner PJ, Gage FH, Casaccia-Bonnefil P, Gu C, Chao M V., Wang S, Barres BA, Geppert M, Sudhof TC, Littleton JT, Bellen HJ, Maximov A, Sudhof TC, Bezprozvanny I, Lemke G, Perrimon N, Bernfield M, Lindahl U, Kusche-Gullberg M, Kjellen L, Hurskainen TL, Hirohata S, Seldin MF, Apte SS, Black RA, White JM, Aravind L, Dixit VM, Koonin E V., Garcia-Meunier P, Etienne-Julan M, Fort P, Piechaczyk M, Bonhomme F, Mansur NR, Meyer-Siegler K, Wurzer JC, Sirover MA, Tatton NA, Kenmochi N, Chen FW, Ioannou YA, Madsen HO, Poulsen K, Dahl O, Clark BF, Hjorth JP, Chambers DM, Peters J, Abbott CM,

Khalyfa A, Carlson BM, Carlson JA, Wang E, Aeschlimann D, Thomazy V, Munroe P, Wu SM, Cheung WF, Frazier D, Stafford DW, Furie B, Kehoe JW, Bertozzi CR, Pawson T, Nash P, Velden AW van der, Thomas AA, Fraser CM, Tettelin H, Brett D, Muller HJ, Kern H, Feinberg AP, Collins CA, Guthrie C, Eddy SR, Wang Q, Khillan J, Gadue P, Nishikura K, Holcik M, Sonenberg N, Korneluk RG, McKinsey TA, Zhang CL, Lu J, Olson EN, Capanna E, Romanini MGM, Smith JM, Charlesworth D, Charlesworth B, Morgan MT, Bailey JE, Maleszka R, Couet HG de, Miklos GL, Miklos GL, Crutchfield JP, Young K, Gell-Mann M, Lloyd S, Barabasi AL, Albert R, Colucci-Guyon E, Ewing B, Green P, Ewing B, Hillier L, Wendl MC, Green P, Lander ES, Waterman MS, Krogh A, Sjölander K, Sjölander K, Bairoch A, Apweiler R, Tatusov RL, Galperin MY, Natale DA, Koonin E V. (2001) The sequence of the human genome. Science 291:1304–51. doi: 10.1126/science.1058040

Wicker T, Mayer KF, Gundlach H, Martis M, Steuernagel B, Scholz U, Simkova H, Kubalakova M, Choulet F, Taudien S, Platzer M, Feuillet C, Fahima T, Budak H, Dolezel J, Keller B, Stein N (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23:1706–1718. 10.1105/tpc.111.086629

Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, Deangelis YM, Reeder NL, Johnstone KR, Leland M, Fieno AM, Begley WM, Sun Y, Lacey MP, Chaudhary T, Keough T, Chu L, Sears R, Yuan B, Dawson TL (2007) Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. Proc Natl Acad Sci U S A 104:18730–5. doi: 10.1073/pnas.0706756104

Zamudio N, Bourc'his D (2010) Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Hered 105:92–104. doi: 10.1038/hdy.2010.53

Zilberman D, Cao X, Jacobsen SE (2003) ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science 299:716–719.doi: 10.1126/science.1079695

Chapter II: Distribution, activity and functional characterization of helitron transposons in *Pleurotus ostreatus*:

This chapter has been published as: Castanera R, Pérez G, López L, Sancho R, Santoyo F, Alfaro M, Gabaldón T, Pisabarro AG, Oguiza JA, Ramírez L (2014) Highly expressed captured genes and cross-kingdom domains present in Helitrons create novel diversity in *Pleurotus ostreatus* and other fungi. BMC Genomics 15:1071. doi: 10.1186/1471-2164-15-1071.

2.1. Introduction

Transposable elements (TEs) are mobile genetic units that impact genome organization and functionality by promoting chromosomal rearrangements and changes in gene structure and expression, among others. Recently, a novel group of Class II DNA TEs called helitrons was detected in *Arabidopsis thaliana* and *Caenorhabditis elegans* by a repeat-based computational analysis (Kapitonov and Jurka 2001). Helitrons are rolling-circle transposons that have been found in plants, protozoans, fungi, cnidarians, insects, worms, fishes, frogs, reptiles and mammals (Poulter et al. 2003; Kapitonov and Jurka 2007; Thomas et al. 2010). These elements are characterized by their 5′TC and 3′CTRR conserved ends as well as a 16- to 20-nucleotide hairpin-forming sequence located approximately 12 nucleotides upstream of the 3′CTRR end (Kapitonov and Jurka 2001). Helitrons lack TIRs, do not generate TSDs upon insertion, and are thought to transpose through a replicative rolling circle mechanism (Kapitonov and Jurka 2001; Feschotte and Wessler 2001) similar to that of bacterial IS91 elements (Toleman et al. 2006). Nevertheless, footprints of helitron somatic excisions have been recently reported in the maize genome, indicating that they may exhibit both replicative and excision-mediated modes of transposition (Li and Dooner 2009). Putative autonomous helitrons contain genes encoding a RepHel protein with a rolling-circle replication initiator (Rep) and a helicase (Hel) domain. Both domains are thought to be essential for transposition. The Rep domain is most likely involved in endonucleolytic DNA breaks during the excision and re-ligation of the transposed DNA (Kapitonov and Jurka 2007). The Hel domain encodes a 5′-3′ DNA helicase in the PIF1/RRM3 family that is highly conserved from yeasts to humans and contributes to the maintenance of genome stability (Boulé and Zakian 2006). When helitrons transpose, they are inserted into AT dinucleotides. One of the most enigmatic features of Helitrons is that during their transposition they can capture, amplify and disperse complete genes and gene fragments by a yet unknown mechanism, playing an important role in the creation of new proteins via exon shuffling and gene duplication (Morgante et al. 2005; Yang and Bennetzen 2009a). According to (Yang and Bennetzen 2009a), most of the genes captured by helitrons in maize are subjected to genetic drift, although 4% of them are subjected to purifying selection and 4% of them to adaptive selection, suggesting that its retention in the genome might be beneficial for the host. In addition, they produce breaks in genetic collinearity, as previously described in maize haplotypes (Lal and Hannah 2005). Helitrons have highly variable lengths (ranging from 202 bp to 35.9 kb in maize) and abundance in eukaryotic genomes. In the fruit fly *Drosophila melanogaster*, helitrons account for 1 to 5% of the total size of the genome (Kapitonov and Jurka 2007), and in mammals such as *Myotis lucifugus* they account for 3% (Pritham and Feschotte 2007). In plants, the contribution of helitrons to the total genome size is variable. In *A. thaliana*, helitrons account for

more than 2% (Kapitonov and Jurka 2001), whereas in *Oryza* the estimations vary from 0.03 in *O. brachyantha* (Zuccolo et al. 2007) to 4% in *O. sativa* (Xiong et al. 2014). In maize, where they have been better characterized (Morgante et al. 2005; Barbaglia et al. 2012), the latest analysis reports the presence of 31.233 helitron copies accounting for 6.6 % of the B73 reference genome (Xiong et al. 2014). In fungi, helitron-like sequences have been identified *in silico* in the genomes of species belonging to the phylum Ascomycota (such as *Aspergillus nidulans, Chaetomium globosum* or *Fusarium oxysporum*) as well as in the zygomycete *Rhizopus oryzae* and the phylum Basidiomycota (such as *Phanerochaete chrysosporium, Coprinopsis cinerea, Laccaria bicolor* or *Puccinia graminis*) (Cultrone et al. 2007; Kapitonov and Jurka 2007; Feschotte et al. 2009; Labbe et al. 2012). However, most of these studies reported on the presence of helitron hits based on similarities to other helitrons in public databases; their enzymatic and structural features were unknown. Thus, we lack a general picture of the structure of fungal helitrons, as well as an understanding of their role in gene capture and their broader genomic impact. *Pleurotus ostreatus* is a white rot basidiomycete that is widely used as a model organism. Recently, the genome of the dikaryotic strain N001 of *P. ostreatus* (which is approximately 34 Mb and organized in 11 chromosomes) was sequenced and annotated and the genome sequences of the monokaryotic strains PC9 and PC15 are available (Riley et al. 2014; Castanera et al. 2016). Sequence analysis of both *P. ostreatus* strains revealed the presence of helitrons in strain-specific genomic locations, as described for different maize haplotypes. With the aim of uncovering new insights into the role of helitrons in the *P. ostreatus* genome as well as the genomes of other ascomycetes and basidiomycetes, we report on: i) their structural features and functional domains, ii) their abundance and occurrence in PC9 and PC15 genomes, and iii) their potential ability to capture, create and express new genes. Finally, we investigate the helitron landscape in *P. ostreatus* and other sequenced fungi to understand their origins and evolution in the fungal kingdom.

2.2. Materials and methods

Structure-based identification of *P. ostreatus* helitrons

The unmasked assembled genomes of the *P. ostreatus* monokaryotic strains PC15 and PC9 were obtained from the MycoCosm database (Grigoriev et al. 2014). The specific web repositories for both genomes are http://genome.jgi-psf.org/PleosPC15_2/PleosPC15_2.home.html (for PC15) and http://genome.jgi-psf.org/PleosPC9_1/PleosPC9_1.home.html (for PC9). Both strains were obtained after de-dikaryotization of the strain N001 (Larraya et al. 1999) and are deposited in the Spanish Type Culture collection (PC9: CECT20311 and PC15: CECT20312). The program HelSearch (Yang and Bennetzen 2009b) was used to analyze the genomic sequences using the eukaryotic consensus 3′- end helitron structure: a minimum of 6 hairpin pairs (two mismatches allowed) located upstream of a 3′ CTRR motif, a 2-4-bp hairpin loop, and 5–8 bp between the hairpin and the 3′CTRR terminal end. The elements detected by HelSearch were classified and aligned into families according to the conservation of their 3′ ends (30 bp with at least 80% identity). The alignment files produced by HelSearch (one per putative family) were manually inspected using MEGA5 (Kumar et al. 2008) to identify the 5′ and 3′ boundaries of each helitron. Elements displaying unclear 3′ boundaries (those without a sharp decrease of similarity in the alignment immediately after the 3′CTRR end) were not used for further analysis. Intact helitrons were defined as elements displaying 5′ and 3′ ends, while truncated elements were defined as those containing an intact 3′ end but not a conserved 5′ end.

Homology-based identification of putative autonomous helitrons

The coordinates of the alignment files produced by HelSearch were to obtain the 5′ upstream regions of each putative helitron end (helend) structure of all the aligned sequences (3,600 bp). The genomic sequences were translated to proteins using the three forward reading frames and subjected to a Batch CD Search (plus and minus strands, $p < 0.01$) (Marchler-Bauer and Bryant 2004) to identify conserved domains. Elements containing Helitron helicase-like (Pfam PF14214) and PIF1-like helicase (Pfam PF05970) domains within the 5′ and 3′ boundaries were considered to be putative autonomous helitrons. Additional *P. ostreatus* helitron-specific helicases were obtained by TBLASTN searches (with a cutoff E-value $<10^{-5}$) using the above mentioned functional domains as queries. Filtered gene models predicted by the JGI and classified as PIF1/DDR3 helicases according to the EuKaryotic Orthologous Groups (KOG) database (Koonin et al. 2004) were also incorporated into the analysis. *P. ostreatus* helitron-specific helicases were aligned using Clustal Omega (Sievers

et al. 2011). The alignments were extended upstream and downstream of the 5′ and 3′ ends to identify the helitron boundaries (Fig 1).

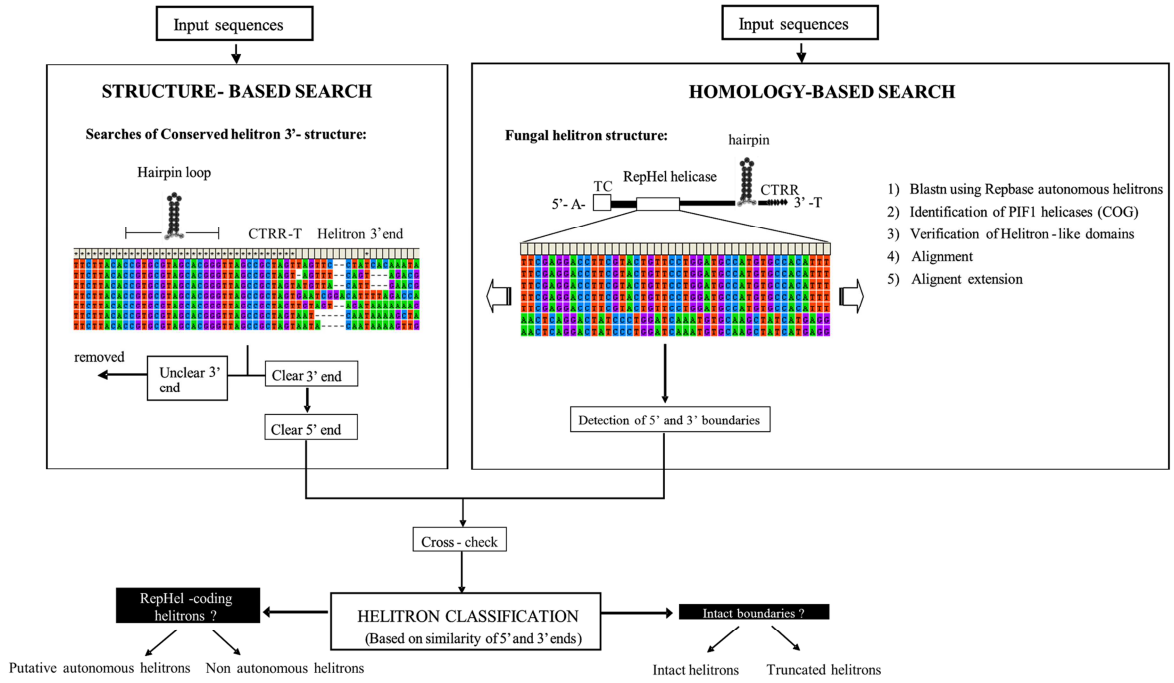

Figure 1. Pipeline for helitron identification and classification in the *P. ostreatus* PC9 and PC15 genomes.

Helitron classification

Elements displaying a nucleotide similarity of 80% or higher in the 30-bp 3′ end were considered to belong to the same family. Elements that met this requirement but had a similarity lower than 80% in the 5′ 30-bp end were classified as a subfamily, according to (Yang and Bennetzen 2009b). Helitrons were named using "HELPO" (**Hel**itron ***P****leurotus* ***o****streatus*) to define the TE class and species, followed by two numbers to define the family and subfamily assignment (i.e., HELPO1.2 belongs to family 1 and subfamily 2). Upright letters are used when referring to families and subfamilies, and italics are used for specific copies (i.e., the HELPO1.1 subfamily vs the *HELPO1.1* element). Putative autonomous elements are shown in uppercase letters, and non-autonomous elements are shown in lowercase letters.

Helitron gene capture

The presence of full-length genes within the boundaries of intact helitrons was analyzed using the JGI genome browser. Predicted gene models (except RepHel helicases) were considered to be captured genes. The presence of these genes in other fungi was analyzed using BLASTX searches of the MycoCosm and NCBI databases (with a cutoff E-value $<10^{-10}$). In addition, BLASTN searches were performed using intact helitrons against *P. ostreatus* assembled scaffolds to find captured gene fragments (hits that were greater than 50 bp and showed more than 95% identity below a cutoff E-value $<10^{-5}$ were considered to be significant). The promoter regions of the captured genes were examined from the start of the RepHel helicase ORF to the start of the captured gene. These regions were subjected to BLASTN searches against the MycoCosm and ViroBlast (Deng et al. 2007) databases (cut-off E-value $<10^{-5}$).

Whole genome alignment.

A whole genome alignment between PC15 and PC9 genomes was performed using the Mercator and MAVID pipeline (Dewey 2007). The adjacent regions of each PC15 helitron shown in Table 1 were analyzed (50 kb upstream and downstream). The locations of every gene placed in these 50kb windows were used to extract individual alignments between PC15 and PC9. The alignments were parsed using Python scripts, and a break of gene collinearity was considered when a gene was present in PC15 but absent in PC9. A gene was considered absent when the alignment length of PC9 was lower than 20% of the length in the PC15 locus, and the frequency of collinearity breaks per every 50kb window was calculated. The same procedure was applied to the full chromosome I, and results were used as a reference estimation of the whole genome. This chromosome was chosen because it is almost fully assembled into a single scaffold in both genomes. The PC15 loci absent in PC9 were used as BlastN query (cutoff E-value < 10-15 and 95% similarity) to check if they were present at a different location. In addition, the same loci were used in a BlastX search (cutoff E-value < 10-5) in the Repbase peptide database (Jurka 2000) to check if they matched to other transposable elements.

RNA-seq data analysis

RNA-seq data from N001 were used to analyze the transcriptional activity of the helitrons and their captured genes. Transcriptome libraries were generated and sequenced by Sistemas Genómicos S.L. (Valencia, Spain) on a SOLiD platform. RNA-seq reads were mapped to the *P. ostreatus* PC15

(assembled into 11 scaffolds) and PC9 (assembled into 572 scaffolds) genome sequences using TopHat (Trapnell et al. 2009), allowing multiple mapping when identical alignment scores where obtained. Transcriptional levels of each helitron were calculated in RPKMs (Reads Per Kilobase per Million mapped reads). The IGV tool (Robinson et al. 2011) was used to analyze the distribution of RNA-seq reads mapping inside the helitron boundaries.

The search for Helitron-like helicases in fungi and other eukaryotes

A TBLASTN search was carried out against the whole fungal MycoCosm database (unmasked assembly scaffolds with a cutoff E-value $<10^{-5}$) (Grigoriev et al. 2014) using the two helitron conserved domains (PF14214 and PF05970) as queries. The results were considered to be an indicator of the presence or absence of putative autonomous helitrons in the different fungal species.

Simultaneously, protein models annotated as DNA helicase PIF1/RRM3 (KOG0987) at the Cluster of Orthologous Groups database were downloaded (2,175 sequences from 284 fungal genomes) and subjected to a Batch Conserved Domain Database Search using a cut-off E-value $<10^{-5}$. Elements carrying the PF14214 and PF05970 domains were kept for further analysis. The eukaryotic putative autonomous helitrons deposited in Repbase (Jurka 2000) (213 sequences) were downloaded, translated to protein sequences using the three forward reading frames and analyzed as mentioned above. Helitron-like helicases from both searches were combined, and those carrying both conserved domains were used for further phylogenetic analysis.

Phylogenetic reconstruction of RepHel helicases

Sequences were aligned using the PhylomeDB pipeline (Huerta-Cepas et al. 2008). In brief, three different alignment algorithms were used: MUSCLE v3.8 (Edgar 2004), MAFFT v6.712b (Katoh et al. 2002), and Kalign (Lassmann and Sonnhammer 2005), in the forward and reverse directions (i.e., using the Head or Tail approach) (Landan and Graur 2007). The six resulting alignments were then combined with M-COFFEE (Wallace et al. 2006) and trimmed with trimAl v1.3 (Capella-Gutierrez et al. 2009) to remove gappy regions and regions that were inconsistent across the reconstructed alignments (with a consistency-score cut-off of 0.1667 and a gap-score cut-off of 0.9). Next, maximum likelihood (ML) trees were reconstructed. First, a tree topology estimated by neighbor joining with BioNJ (Gascuel 1997) was used to infer the likelihood of seven different evolutionary models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff). The best model fitting data as determined by the AIC (Akaike's Information Criterion) were used to derive ML trees using phyML v 3.0 with four rate categories and inferring invariant positions from the data

(Guindon et al. 2010). Branch support was computed using an a LRT (approximate likelihood ratio test) based on a chi-square distribution. The tree figures were produced using ETE v2 (Huerta-Cepas et al. 2010).

Strains and culture conditions

The *P. ostreatus* monokaryotic strains PC15 and PC9 and the dikaryotic strain N001 were grown in triplicate on a submerged SMY medium (10 g/L saccharose, 10 g/L malt extract, 4 g/L yeast extract). Shaking cultures (130 rpm) were kept in the dark at 24 °C for 8 days.

Nucleic acid extraction and reverse transcription

Total RNA was extracted from ~200 mg of deep frozen tissue using the Fungal RNA E.Z.N.A. Kit (Omega Bio-Tek, Norcross, GA) and treated with 1 U of RQ1 DNase (Promega, Madison, WI) per µg of RNA. The RNA integrity was estimated using denaturing electrophoresis on 1% (w/v) agarose gels. The nucleic acid concentrations were measured with a Qubit 2.0 fluorometer (Life Technologies), and the purity of the total RNA was estimated using the 260/280 nm absorbance ratio on a NanoDrop™ 2000 (Thermo Scientific) machine. The total RNA (225 ng) was reverse-transcribed into cDNA in a 20-µl volume using the iScript cDNA Synthesis kit (Bio-Rad, Alcobendas, Spain).

Real-time PCR

The amplifications were performed using a Bio-Rad CFX96 thermal cycler. SYBR green fluorescent dye was used to detect the product amplification. Each reaction was set to a final volume of 20 µl and contained 1X IQ SYBR green Supermix from Bio-Rad, 300 nM forward and reverse primers (Supplementary information, Table S1), and 1 µl of a 1:20 dilution of RT product in nuclease-free water. The amplification program consisted of 5 min at 95 °C, 40 cycles of 15 s at 95 °C and 30 s at 60 °C, followed by 1 min at 95 °C, 1 min at 65 °C with a final melting curve with increments of 0.5 °C every 5 s in a linear gradient of 65 to 95 °C. High-temperature fluorescence acquisition (72 °C) was performed to eliminate the impact of the PCR artifacts in cDNA quantification, and the absence of these artifacts was confirmed by a melting-curve analysis. A baseline correction and crossing-point (Cp) acquisition were performed using Bio-Rad's CFXManager. The reactions were performed in triplicate in 96-well microtiter plates. NRTs (non-retrotranscribed controls) and NTCs (no-template controls) were included for each primer set. The amplification efficiencies were sample-estimated by a linear regression from a window-of-linearity set in the exponential phase of the fluorescence history plotted in log scale using the LinReg tool

(Ramakers et al. 2003). Raw Cp values were efficiency corrected, and any signal of genomic DNA background was removed using GENEX (http://www.multid.se.). The transcription level of each gene of interest (GOI) was calculated as a relative quantity (RQ, equation 1) using *pep* as an internal standard.

Equation 1: $$RQ = 2^{-(Cp_{GOI} - Cp_{PEP})}$$

2.3. Results

Pleurotus ostreatus helitrons

We designed a pipeline for helitron identification in *P. ostreatus* (Fig 1) starting with a structure-based approach using HelSearch, which scans the genome looking for sequences compatible with helitron 3'-end conserved structure. This approach yielded 11 and 9 putative helitron families in the PC15 and PC9 genomes, respectively (Supplementary information: Table S2). Our subsequent homology-based approach uncovered another putative helitron family that could not be detected by the first method. After a manual curation of the alignments and the removal of false positives, we obtained two verified helitron families named HELPO1 and HELPO2. Both families contain most of the structural and enzymatic features described earlier in plant/animal helitrons such as AT insertion specificity, T[C/G]-5′ and CTRR-3′ ends (CTTG in the case of HELPO2), the presence of a subterminal palindromic hairpin, and a rolling-circle replication initiator as well as a helicase domain in a common ORF (Fig 2). Based on the similarity of the 5′ and 3′ boundaries (see Materials and Methods), helitrons of the HELPO1 family can be further classified into three subfamilies: HELPO1.1, HELPO1.2 and HELPO1.3, with elements ranging from 1.5 to 13.7 kb length (Fig 2.A). HELPO2 contained elements varying from 3.9 to 10.6 kb in length. Both the HELPO1 and HELPO2 families contain putative autonomous elements, however, the HELPO1 family is the only one carrying intact non-autonomous copies, all of them belonging to subfamily HELPO1.3 (Fig 2). The flanking regions of the helitron insertion sites (50 bp) are AT-rich (AT content of 57%), whereas the AT content in the internal regions is similar to that of the whole genome (49%). The putative autonomous elements of the HELPO1 and HELPO2 families carry an ORF encoding a RepHel helicase of approximately 1,400 aa. The protein contains three motifs defining the rep domain (Kapitonov and Jurka 2007) as well as six conserved motifs present in members of the SF1 helicase superfamily described in other helitrons (Supplementary information: Fig S1) (Pritham and Feschotte 2007; Han et al. 2013) and necessary for replication and DNA unwinding. Using a maximum likelihood approach, we clustered the RepHel helicases into three groups (Supplementary information, Fig S1), where the *HELPO1* and *HELPO2* proteins grouped separately. Interestingly, the third group lacks the rolling-circle replication initiator but ferries some of the helicase domains. Apparently, these helicases do not belong to a specific helitron family. It should be pointed out that putative HELPO1 and HELPO2 autonomous elements share about 60-70% similarity to Helitron 1_SLL_1p of *Serpula lacrymans (Eastwood et al. 2011)* and Helitron2_Ppa_1p of *Physcomytrella* (Jurka 2010), but only in the regions corresponding to the Helitron helicase-like domain (Pfam PF14214) and the PIF1-like helicase domain (Pfam PF05970)

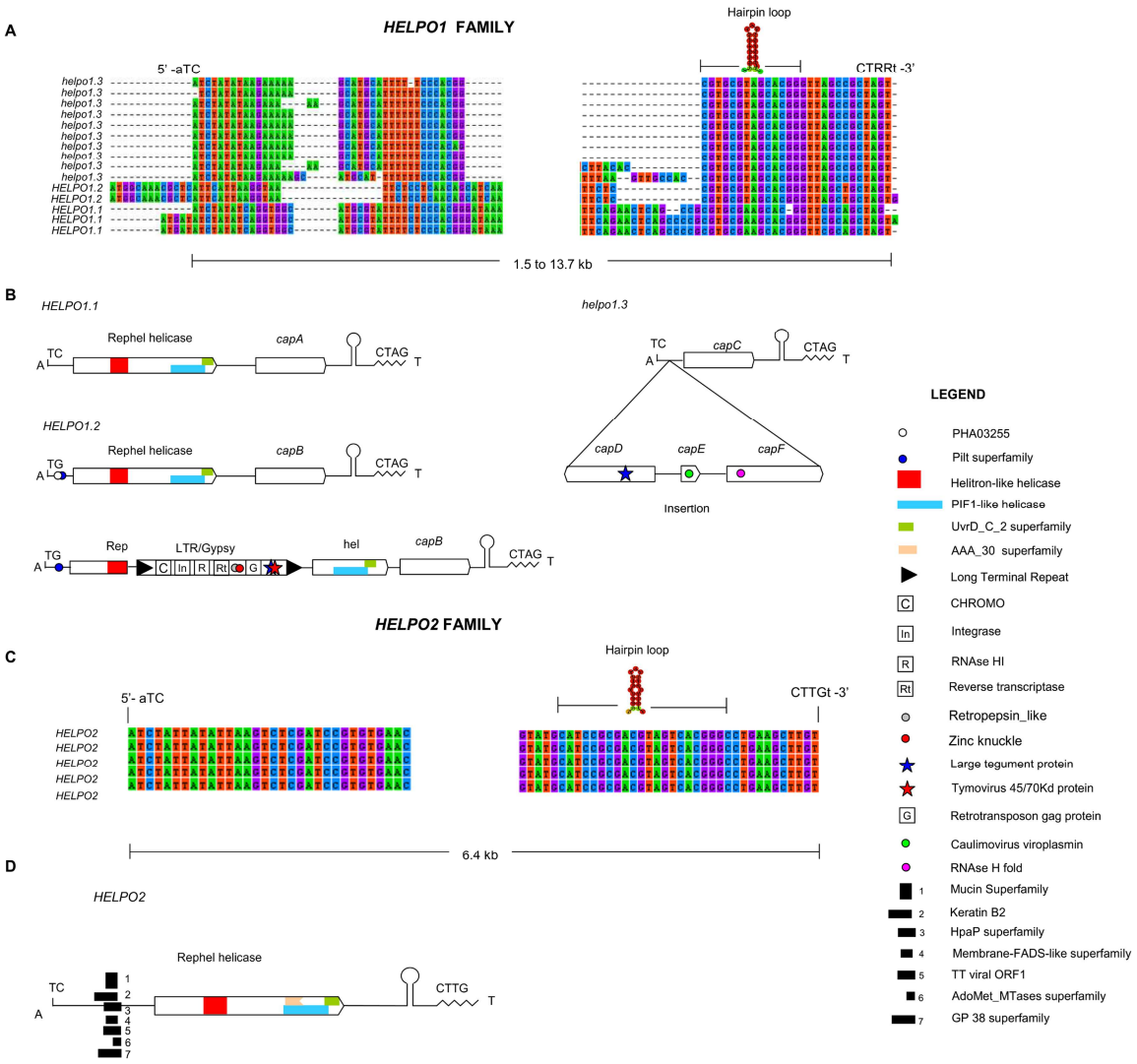

Figure 2. Structural and enzymatic features of the *P. ostreatus* helitron families. Alignments of the 5′ and 3′ boundaries of the helitron families HELPO1 (A) and HELPO2 (C). Schematic representation of the structural hallmarks, coding features and conserved domains (CDD, cutoff E-value <0.01) of the different elements belonging to the HELPO1 (B) and HELPO2 (D) families

Helitron abundance in the *P. ostreatus* PC15 and PC9 homologous genomes

A total of 37 validated helitrons in the HELPO1 and HELPO2 families were detected in the PC15 strain (Table 1), accounting for 0.35% of the total genome size. Among these helitrons, 19 were intact elements, and 11 out of the 19 were full-length putative autonomous elements. The remaining elements were truncated copies. In the PC9 genome, 10 helitrons accounting for 0.05% of its

genome were found, of which only five could be mapped to the corresponding PC15 scaffolds (Fig 3A). Five elements showed intact 5′ and 3′ boundaries, one was putative autonomous (*HELPO1.1*), and the rest were truncated elements. Helitron length polymorphisms were observed in some of the elements. PC15 HELPO1.2 subfamily showed two elements of different lengths. The shortest element (7.1 kb) was located on chromosome XI, and the largest (13.3 kb) was located on chromosome I. The HELPO1.3 subfamily was the only subfamily with non-autonomous elements at identical positions in both genomes**.** In this sense, it should be noted that the large *helpo1.3* copy appeared as an allele of the short copy on chromosome I. Copies of the short *helpo1.3* copy were also found on chromosome II. In PC15, helitrons were found in ten out of eleven chromosomes. Seven chromosomes carried helitrons from both families, while three (chromosomes II, VI and X) carried helitrons from only a single family. Chromosomes I, VII and VIII carried the highest number of helitrons. Clusters of helitrons were present in the regions of chromosomes I and VII (Fig 3A). Breaks in gene collinearity between PC15 and PC9 were observed in 66% of the helitron containing regions (except in the genome regions described above), as shown in Fig 3. The analysis of 44 regions of 50 kb adjacent to HELPO1 and HELPO2 helitrons revealed that the frequency of collinearity breaks in these regions was 1.86 every 50 kb, while the frequency in the whole chromosome I was 1.25 breaks every 50 kb. According to our results, 40% of the PC9 missing counterparts were present in a different location, while 22% corresponded to other transposable elements, mainly LTR/Gypsy, DNA/PIF-Harbinger and DNA/CMC-EnSpm. In chromosome VII, the two *HELPO1.1* copies showed 99.7% similarity. One of the copies was inserted into the left 576-bp inverted repeat found in a 37.2-kb region present on a chromosome of PC15 but was absent in the PC9 genome (Fig 3C). This region was also found close to the telomere in chromosome XI of PC15 and carried 14 predicted genes, including a CACTA transposase.

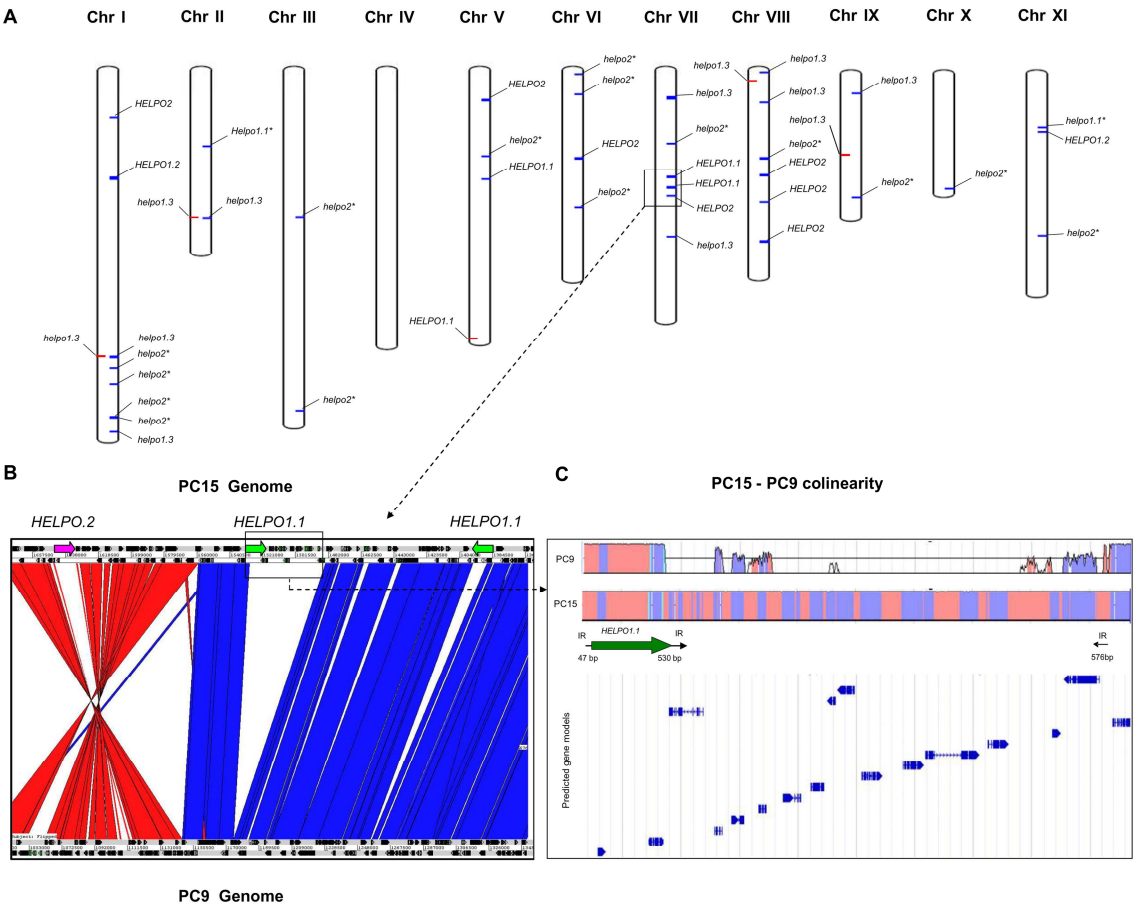

Figure 3 Helitrons break the synteny between the *P. ostreatus* PC15 and PC9 genomes. The distribution of helitrons in the chromosomes of the dikaryotic strain N001 is shown in A (PC15 elements are shown in blue and in PC9 elements are shown in red). Truncated elements are marked with a '*'. An ACT (Carver et al. 2005) comparison of the squared region between PC15 and PC9 is shown in B. The lack of gene collinearity between PC9 and PC15 in the squared region of chromosome VII is shown in C (coordinates: 1,528,715-1,479,715). In the synteny plot, coding regions are represented in purple and inter-genic regions in pink. Arrows labeled IR represent the inverted repeats found in a 37.2 kb region duplicated in PC15 and absent in PC9 genome. Blue arrows underneath synteny plot represent predicted genes.

Table 1. Summary of the helitron characteristics in the *P. ostreatus* dikaryotic strain N001

Name/ID	Genome	Scaffold*	Start (bp)	End (bp)	Orientation	Size (Kbp)	Autonomous	Captured genes	Intact	KPKM
HELPO1.1	PC15	5	1418,337	1425442	-	7,1	*	*cap A2*	*	38.64
HELPO1.1	PC15	7	1521533	1528715	-	7,2	*	*cap A*	*	13.64
HELPO1.1	PC15	7	1387822	1395008	+	7,2	*	*cap A*	*	16.05
HELPO1.1	PC15	11	702113	708712	+	6,6	*	*cap A*		0
HELPO1.2	PC15	1	1404200	1417948	+	13,7	*	*cap B*	*	1.38
HELPO1.2	PC15	11	756984	764123	+	7,1	*	*cap B*	*	2.41
helpo1.3	PC15	1	3742549	3754728	+	12,2		*cap C, Cap D, cap E, Cap F*		
helpo1.3	PC15	1	4537580	4539124	-	1,5		*cap C*	*	5.43
helpo1.3	PC15	2	1934871	1936414	-	1,5		*cap C*	*	7.69
helpo1.3	PC15	7	380915	382459	+	1,5		*cap C*	*	6.33
helpo1.3	PC15	7	2181803	2183334	+	1,5		*cap C*	*	5.04
helpo1.3	PC15	8	26555	28057	-	1,5		*cap C*	*	3.29
helpo1.3	PC15	8	423643	425136	-	1,5		*cap C*	*	5.45
helpo 1.3	PC15	9	258831	261375	-	1,5		*cap C*	*	6.23
HELPO 2	PC15	1	619761	626150	+	6,4	*		*	0.18
HELPO 2	PC15	5	387607	398218	-	10,6	*			0.14
HELPO 2	PC15	6	1150050	1156438	-	6,4	*		*	0.2
HELPO 2	PC15	7	1635256	1641644	-	6,4	*		*	0.19
HELPO 2	PC15	8	1367660	1374048	+	6,4	*		*	0.19
HELPO 2	PC15	8	2234922	2241310	-	6,4	*		*	0.18
HELPO 2	PC15	8	1722302	1726240	+	3,9	*			0.03
HELPO 2	PC15	11	2114358	2115721	+	1,4	*			0
154430	PC15	2	1672984	1678495	-	5,5				0
1035322	PC15	2	1778564	1779100	-	0,5				0
1044620	PC15	7	2753773	2756802	+	3				0
1078941	PC15	8	2474562	2480319	+	5,8				0
1078947	PC15	8	2505293	2510116	-	4,8				0
1079561	PC15	10	1356533	1360752	-	4,2	*			1
HELPO 1.1	PC9	115	1	7176	-	7,2	*	*cap A*	*	17.59
HELPO 1.1	PC9	91	1	560	+	0,6		*cap A*		
HELPO 1.2	PC9	366	1	3061	+	3,1	*	*		0.25
helpo 1.3	PC9	7	1079490	1082511	+	3		*		2.62
helpo 1.3	PC9	44	2611	4188	-	1,6		*	*	6.41
helpo 1.3	PC9	142	1	533	+	0,5		*		5.37
helpo 1.3	PC9	360	2475	3115	+	0,7		*		5.65
helpo 1.3	PC9	375	1	440	-	0,4		*		3.64
helpo 1.3	PC9	478	373	1917	-	1,5			*	9.27
HELP 02	PC9	440	1	2592	-	2,6		*		0
48294	PC9	2	447393	449295	+	1,9				3.07
51890	PC9	3	2765812	2767353	+	1,5				0
52065	PC9	3	5913	9386	+	3,5				0

* Indicates that the element fits the description shown in the header

Helitron captured genes

The helitrons of the HELPO1 family show a high tendency for gene acquisition/creation, as every intact copy carried gene-like sequences (Fig 2B, Table 1). By contrast, members of HELPO2 only contained the RepHel helicase. In PC15, putative autonomous elements of the HELPO1 family carried from one to four captured genes (*cap*) downstream of the RepHel helicase. The captured genes of the HELPO1.1 subfamily were named *capA*, those from HELPO1.2 were named *capB*, and those from HELPO1.3 were named *capC*, *capD*, *capE* and *capF* (Fig 2B). The captured gene of the HELPO1.1 copy on chromosome V was named *capA2* instead *of capA* due to its low similarity to the other *capA* genes (45%, Supplementary information: Table S3) carried by the helitrons on chromosome VII. Chromosome XI harbors a *capB* gene in a *HELPO1.2* element. Interestingly, an extra copy of the HELPO1.2 subfamily carrying (apart from the *capB* gene) a Gypsy LTR-retrotransposon was found on chromosome I (>70% similarity of nucleotide sequence to Gypsy-8_CCO-I of *Coprinopsis cinerea deposited in Repbase*). The Gypsy element was inserted in the second reverse reading frame, breaking the RepHel helicase ORF (Fig 2B). Several copies of this retroelement were found in chromosomes I, III, V, IX and XI of PC15. The genes carried by *HELPO1* helitrons can be classified based on their conserved domains as retrotransposon/viral genes or as genes of unknown function.

Retrotransposon/viral genes

An analysis of the conserved domains showed significant hits (CDD, cutoff E-value <0.01) in a *HELPO1.2* copy harboring LTR/Gypsy and in a *helpo1.3* copy, both present on chromosome 1 (Fig 2B). The *HELPO1.2* copy on chromosome I carried viral and retrotransposon domains in addition to helitron motifs (Supplementary information: Table S4). BLASTN searches performed on PC15 filtered model genes using intact helitrons as queries showed that this *HELPO1.2* was the only helitron harboring plant and animal re-arranged retroviral genes shuttled by a retroelement. The largest *helpo1.3* copy on chromosome I was 10.7 kb longer than the mean of the lengths of the other *helpo1.3* copies in the *P. ostreatus* genome (12.2 kb vs. 1.5 kb, Fig 2B and Fig 4), and it bore a small expressed region without a predicted gene model (the *capC* gene) as well as three predicted genes (*capD*, *capE* and *capF*). The *capD* gene contains a domain present in the large tegument protein UL36 of the herpes virus (PHA03247), *capE* carries a Caulimovirus viroplasmin (pfam01693), and *capF* carries a predicted nuclease (RNAse H L fold, COG4328). All of the *cap* genes described above are present exclusively within helitrons and do not have additional copies outside helitron boundaries.

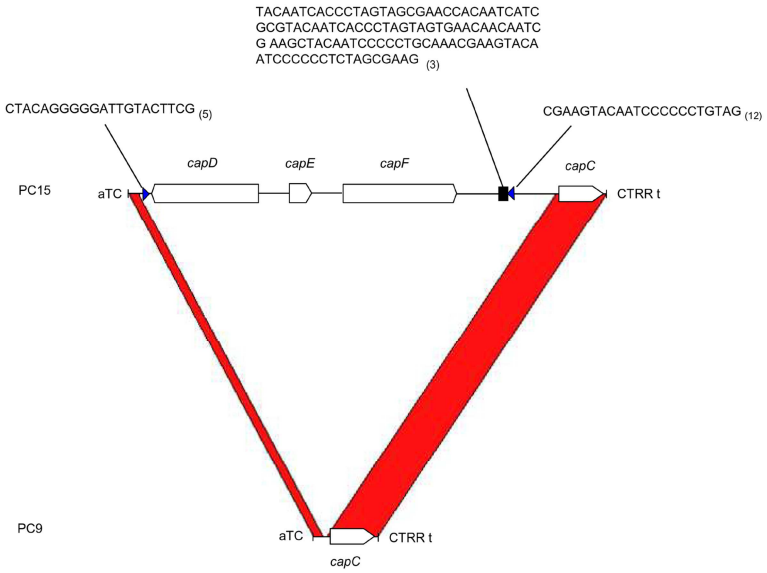

Figure 4. Helitron length polymorphisms in allelic copies of the HELPO1.3 subfamily. Regions in red are highly conserved. Blue triangles represent inverted repeats, and the black square represents a satellite sequence (the number of repeats is shown in parentheses). Empty arrows represent predicted ORFs.

Genes of unknown function

capA, *capA2*, *capB* and *capC*, did not bear conserved domains. A BLASTX query of the entire MycoCosm database (cutoff E-value $<10^{-10}$) revealed that the *capA*, *capA2* and *capC* genes were novel *P. ostreatus*-specific fungal genes, while *capB* yielded significant hits for proteins of unknown function that are present in a few species of Basidiomycetes: *Armillaria mellea* (ID: 8292), *Dendrothele bispora* (ID: 811331), *Fibulorhizoctonia* sp. (ID: 941557), *Schizophyllum commune* Loenen (ID: 271731), and *Suillus brevipes* (ID: 956931). With the exception of *A. mellea* (because the gene was at the end of the scaffold), all of the species carried the RepHel helicase in the same orientation as the *P. ostreatus* helitron *HELPO1.2*, as evidence of the patchy distribution of this helitron subfamily in the phylum Basidiomycota. In addition, no hits for any promoter transcription factor motifs were found in BLAST searches against fungal (MycoCosm) and Viral (viroBlast) databases (cutoff E-value $<10^{-5}$). All these genes of unknown function were further analyzed with Phyre2 (Kelley et al. 2015) to predict their protein structure and perform PSI-BLAST (Position-Specific Iterated BLAST). Using this sensitive approach we could detect the presence of a

Ribonuclease H-like motif in *capA* genes, although with low confidence (60-61%) and alignment coverage (13%).

Transcription

The transcriptional profiles of 30 helitrons and 10 truncated RepHel helicases from the *P. ostreatus* PC15 and PC9 genomes were investigated in solid SMY cultures using RNA-seq in the dikaryotic strain N001 (Fig 5, Table 1). An analysis of RNA-seq reads using IGV yielded different profiles for the members of different families and subfamilies (Fig 5). In most cases, the RNA-seq reads did not fit with the predicted gene models, although we also found RNA-seq reads that mapped to regions with no annotated models (ie, *capC).* Helitrons in the HELPO1 family showed higher levels of transcription (based on the RPKM values of the entire helitron, including the RepHel helicase and the captured genes), in comparison with the elements belonging to the HELPO2 family. The truncated PIF1 helicases showed no transcriptional activity, with the exception of helicase ID 1079561 (on chromosome X). The HELPO1.1 subfamily displayed very high levels of expression (up to 38.64 RPKM) compared with the HELPO1.3 (maximum of 7.69 RPKM) and HELPO1.2 members (maximum of 2.41 RPKM). RT-qPCR experiments were performed using mRNA from the strains PC9, PC15 and N001 grown in submerged cultures to analyze the expression of the RepHel helicases and captured genes independently. For RepHel helicases, similar relative profiles were observed in the three strains, although the ranges of the transcriptional levels were different (Fig 5F). The RepHel helicase of HELPO1.2 was frequently the most highly expressed (0.31, 20.9 and 6.9 RQs in PC9, PC15 and N001, respectively). HELPO1.1 RepHel showed much lower expression levels (0.25, 1.8 and 0.2 RQs in PC9, PC15 and N001, respectively) and HELPO2 showed no expression in N001 and PC9 (0, 2.6 and 0 RQs in PC9, PC15 and N001, respectively).

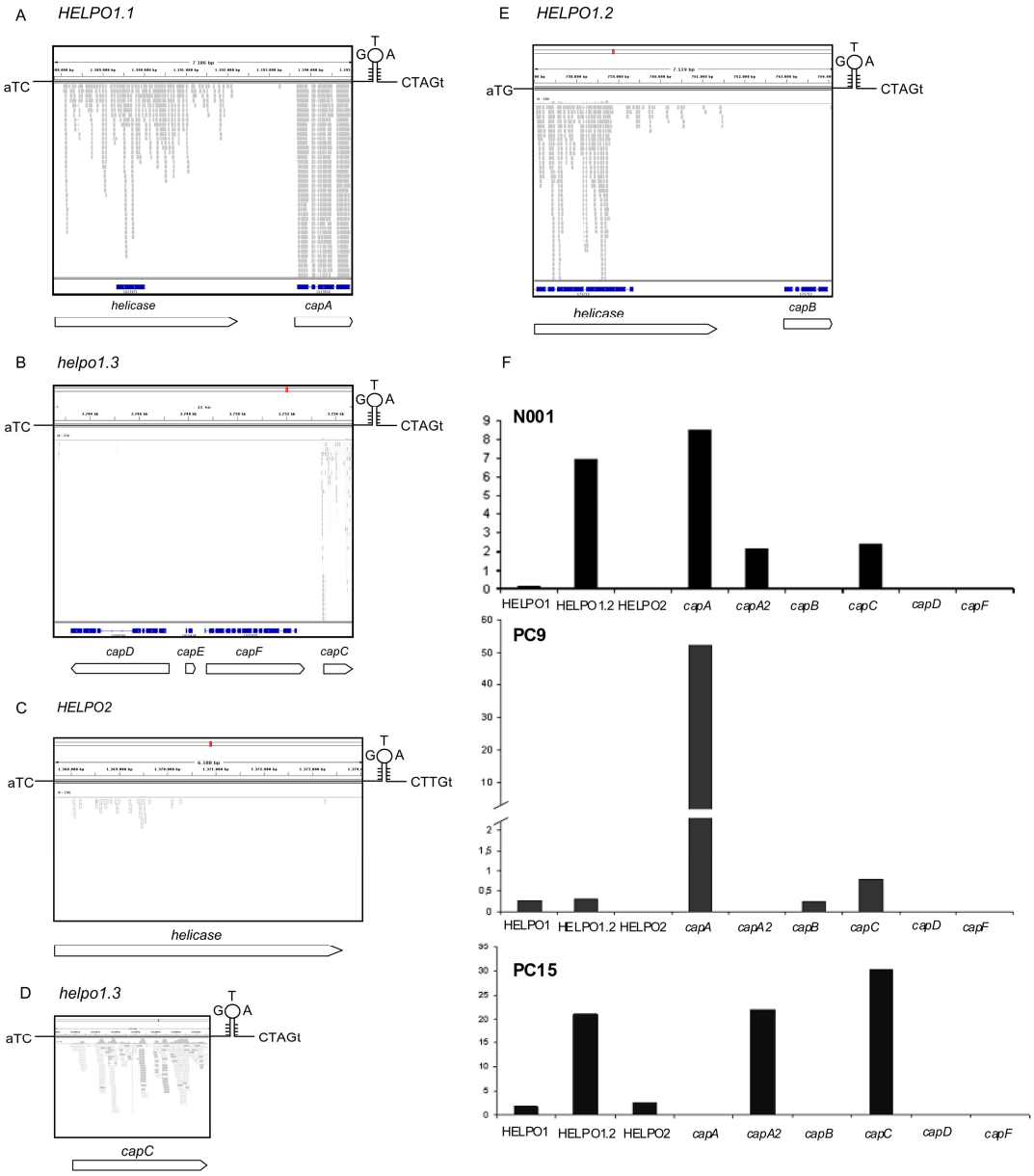

Figure 5. Transcriptional profiles of helitron-specific helicases and captured genes. Five representative RNA-seq profiles of the helitron families and subfamilies (A to E). The gene models predicted by JGI annotation pipeline are shown in blue. Empty arrows represent manually annotated features. The expression of the N001, PC9 and PC15 RepHel helicases and captured genes by RT-qPCR is shown in F. The Y axis of F represents the expression (RQ) relative to the reference gene pep.

Virus-like captured genes carried by LTR/Gypsy did not show transcription in any strain, and genes of unknown function, such as *capA, capA2*, *capB* and *capC*, showed a strain-specific expression profile. RT-qPCR experiments performed with PC9 showed that *capA* was the most highly expressed gene (52.3 RQs) whereas much lower (*capB* and *capC*) or no transcription (*capA2*) was observed for the rest of these genes. In PC15, the highest expression values corresponded to *capC* and *capA2* (30.4 and 22.1 RQs). In N001, c*apA* was the most highly expressed gene, followed by *capC* and *capA2* (8.51, 2.4 and 2.2 RQs) (Fig 5F). Clear differences were observed between the *capA and capA2* transcription profiles using RT-qPCR. Because the primers were designed to amplify more than one gene with the exception of *capA2*, *capD* and *capF* (Supplementary information: Table S1), the transcription levels obtained were the result of the contribution of every RepHel helicase and captured gene from each helitron family.

Differential expansion of the helitron-specific helicases in other fungi

TBLASTN homology-based searches were carried out on the entire MycoCosm database (as of January 2014) using the Helitron helicase-like (PF14214, 182 aa) and PIF1-like helicase (PF05970, 362 aa) domains as queries. The search yielded 1,311 and 1,645 significant hits in 149 genomes (cutoff E-value $<10^{-5}$) to the Helitron helicase-like and PIF1-like helicase domains, respectively. The results were used to analyze the expansion of helitron-specific RepHel helicases in fungal phyla. We found a clear difference in the occurrence of helitron-like helicases in the Ascomycetes and Basidiomycetes classes. While 87% of the genomes of the basidiomycetes analyzed contained RepHel proteins, only 30% of the ascomycetes contained RepHel proteins. This difference is even more striking when we consider that the ascomycetes group comprised a larger number of analyzed genomes. Interestingly, the correlation of the presence of both domains was very high (r =0.91) in fungi.

Phylogenetic reconstruction of eukaryotic RepHel helicases

To investigate the evolutionary relationships of the fungal helitrons identified as well as those from other eukaryotic genomes, we reconstructed molecular phylogenies of the PIF1-like helicase and Helitron helicase-like domains. An initial dataset containing 2,175 PIF1-like helicases from 284 fungal genomes (JGI filtered models) and 213 putative autonomous elements obtained from Repbase (including plants, animals and fungi) was used to uncover new insights into the helitron distribution in the eukaryotic domain. A total of 672 sequences bore the PIF1-like helicase domain, 416 carried the Helitron helicase-like domain, and 125 sequences displayed both domains. After removing duplicated copies, the two functional domains of the remaining sequences were extracted,

aligned and subjected for phylogenetic analyses. Both analyses depicted a similar scenario - fungal helitrons were not monophyletic, but rather they appeared in at least four different clades interspersed among metazoan and plant helitrons (Fig 6). In addition, within each fungal clade, the different fungal phyla (e.g., ascomycetes, basidiomycetes) appeared mixed.

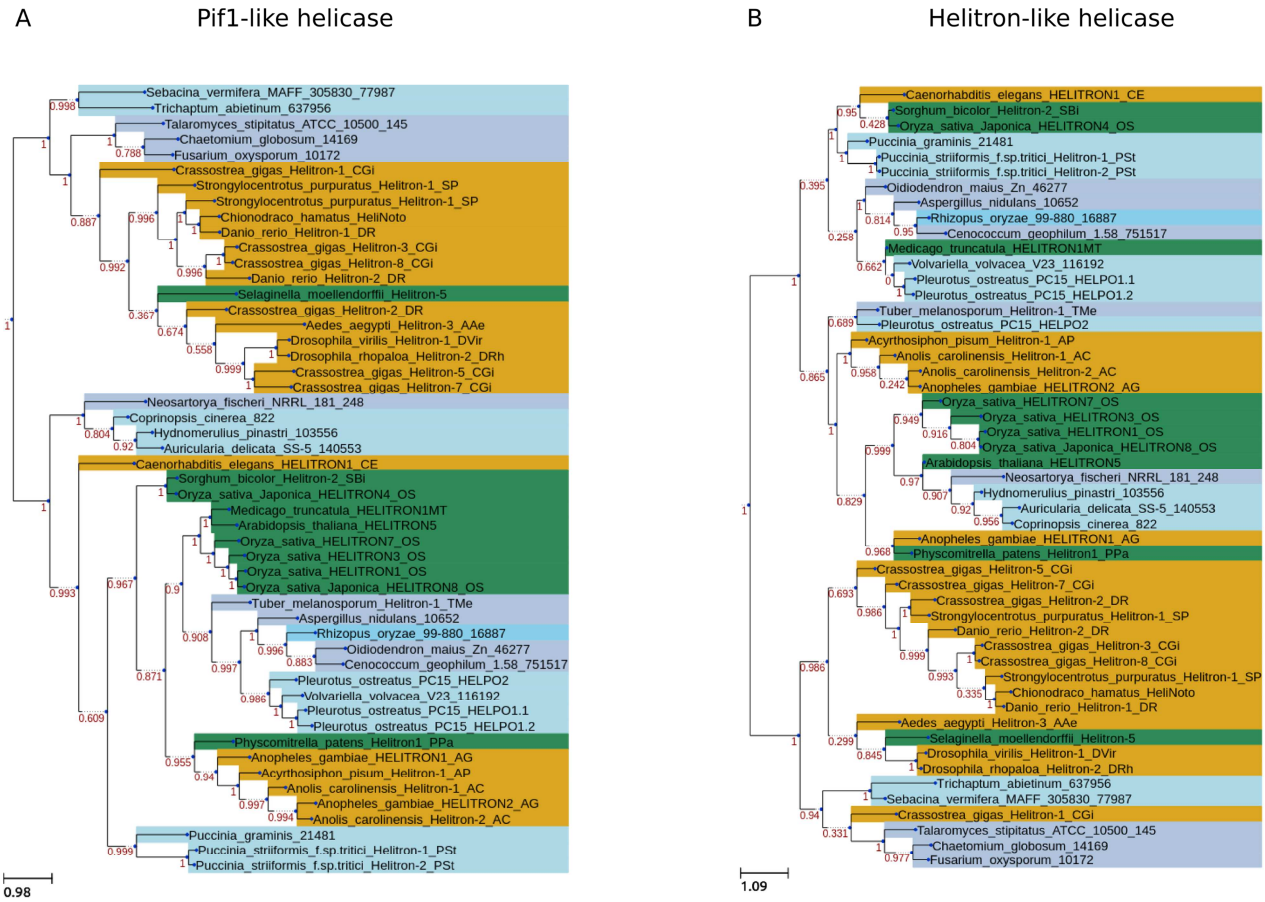

Figure 6. Phylogenetic reconstruction of the eukaryotic Pif1-like helicase domain and Helitron-like helicase domain. Green represents helitrons from the Plant kingdom, yellow from the Animal kingdom, and blue from the fungal kingdom. Light blue represents the phylum Basidiomycota and dark blue represents the phylum Ascomycota.

2.4. Discussion

Previous studies have shown that helitron transposons are widespread in eukaryotic genomes (Kapitonov and Jurka 2007; Pritham and Feschotte 2007; Yang and Bennetzen 2009a). Their structural and enzymatic features have been analyzed in depth in plants and animals using computational analyses, uncovering a canonical structure that is widely conserved among the elements in both kingdoms. Several tools and pipelines have been published for analyzing helitrons in a diverse range of eukaryotic genomes (Du et al. 2009; Yang and Bennetzen 2009b; Han et al. 2013). These approaches rely on either homology-based searches of previously known helitrons or structure-based searches of unique helitron features such as the conserved 3′- terminus. However, fungal helitron-like sequences can lack intact boundaries (Kapitonov and Jurka 2007). This characteristic impedes helitron identification using structure-based searches. In *P. ostreatus*, we show that both the structural and coding features (the Rep and Helicase domains) are present and highly conserved with those present in other helitrons in different kingdoms. Nevertheless, the slight variation in the 3′-terminus of HELPO2 elements makes them undetectable by HelSearch. This situation necessitates combining homology-based searches, structure-based approaches and manual curation for fungal helitron searches, as described in this study. In terms of relative abundance, the helitron content of *P. ostreatus* is similar to that of other basidiomycetes (0 to 0.5% of their genome size, (Eastwood et al. 2011)). We found that genome assemblies of poor quality (ie, with high content of gaps and low L50 values) critically impacted helitron searches, leading to uncertainty in the quantification of helitron content. The *P. ostreatus* PC15 genome sequence was assembled into 11 scaffolds, which fit with the 11 known linkage groups (Larraya et al. 2000). However, the PC15 scaffolds were not used as templates for the PC9 assembly because our goal was to analyze the effect of helitrons and other TEs in breaking synteny and the consequences of hemizygous regions with respect to *P. ostreatus* mushroom yield and enzyme expression. Thus, we found the estimation of helitron abundance for the PC15 genome to be more accurate than the PC9 genome because PC9 is assembled into 572 scaffolds, most of which are very small in size. *P. ostreatus* helitrons insert precisely between A and T nucleotides, and tend to land in AT-rich genome regions as described in maize helitrons (Yang and Bennetzen 2009a). In *P. ostreatus*, approximately half of the helitrons were found in retrotransposon-rich regions. This phenomenon is more pronounced in the *helpo1.3* and *HELPO2* elements because they are more abundant. A high percentage of HELPO1 helitrons were putative autonomous elements carrying captured genes inside their boundaries compared with HELPO2. The similarity between the elements belonging to different families and subfamilies (approximately 40% between HELPO1 and HELPO2, and

approximately 60% between HELPO1.1 and HELPO1.2, Supplementary information: Table S5) strongly suggests that helitron vertical diversification has occurred. However, recent amplification events are not excluded because both the HELPO1 and HELPO2 families contain young elements (i.e., HELPO1.3 and HELPO2 display elements with 99-100% similarity). Notably, the short copy of *helpo1.3* (1.5 kb) occurs frequently in the *Pleurotus* genome compared with the large one (12.2 kb, present only once). The long copy contains internal complementary repeats flanking capD and capF genes. These sequences may have promoted an intrachromosomal rearrangement mediated by the formation of a loop that contains the captured lost genes *capD*, *capE* and *capF*. The short copy of *helpo1.3* would then bear only *capC*, which is later amplified. Alternatively, these three unknown genes may be remnants of an ancient insertion of a virus or a DNA transposon inside *helpo1.3* (ie, a nested transposon). The presence of an RNAse H fold domain (COG4328, transposase-like) in *capF* gives strength to this hypothesis. In the *Pleurotus* genome, the mobilization of LTR/Gypsy elements and their insertion into helitrons creates chimeric elements. For example, a LTR/Gypsy element present in several Basidiomycetes genomes was found in an opposite orientation breaking the RepHel helicase ORF of a *HELPO1.2* element in *P. ostreatus* PC15. This finding supports an insertion rather than a capture of the LTR/Gypsy element by a helitron. This result greatly differs from that found in plants and animals, where helitrons frequently capture gene fragments from their hosts (Pritham and Feschotte 2007; Yang and Bennetzen 2009a; Fu et al. 2013). In this regards, previous studies by found that chimeric elements formed by helitrons and other TEs are rare in eukaryotic genomes (Gao et al. 2012).

Helitron-mediated amplification and expression of captured genes

Pleurotus helitrons contain a subterminal hairpin and a well-conserved 3′-CT[A/T)G end, and do not generate target site duplications in agreement to what was previously described for other eukaryotes (Kapitonov and Jurka 2001; Yang and Bennetzen 2009a). The conservation of the 3′-end structure in helitrons from highly divergent species (i.e., fungi and plants) suggests that the 3′-end structure plays an important role in transposition. Earlier studies have hypothesized that this structure could serve as a terminator transposition signal. In this sense, the read-through-model-1 (RTM1) (Feschotte and Wessler 2001) proposes that a malfunction of this RC terminator may lead to the acquisition of genes or gene fragments adjacent to the 3′ helitron end. The location of captured genes downstream of the RepHel helicase (i.e., *capA* and *capB*) fits with the RTM1 model of gene capturing through new 3′-end acquisition, although there were no clear intermediate RC terminators representing ancient helitron-ends. This could be due to the deletion of the 3′ terminus during transposition or due to sequence degeneration. In fact, the RC terminator in the new transposon would be formed *de novo* by a terminator-like signal in the surrounding location, as

described in the capture of a fragment of the xanthine α-ketoglutarate-dependent dioxygenase gene by a non-autonomous *Helitron-N1_AN* from *A. nidulans* (Cultrone et al. 2007). In plants and animals, helitrons contain genes captured from their hosts (Pritham and Feschotte 2007; Fu et al. 2013). In *P. ostreatus*, the fact that there are very few significant BLAST hits in databases using *capB* as a query in addition to the absence of hits using the other *cap* genes as queries indicates that *cap genes* are either novel structures created by shuffling DNA sequences from diverse origins or the result of a gene capture in a host other than fungi whose sequence is still not available. The difference found in the gene capturing frequencies of the HELPO1 (high frequency) and HELPO2 (no captured genes) families, as well as the scarce and patchy distribution of some of these genes in the fungal phylogeny, gives strength to the hypothesis of an ancient capture in a previous host. The architecture of the non-autonomous copy of the HELPO1.3 family that carries four predicted genes (Fig 2, Fig 4) fits with the filler DNA model (Kapitonov and Jurka 2007) in which the captured regions are acquired by the machinery responsible for the non-homologous repair of double-stranded DNA breaks. A similar mechanism was described for the integration of viruses in chicken cells (Bill and Summers 2004). Recently, due to the increasing number of whole genome sequencing projects and bioinformatics analysis tools available, a large body of literature has been reported regarding virus integration into eukaryotic genomes (endogenous viral elements, EVE) and their roles in their hosts (Katzourakis and Gifford 2010; Feschotte and Gilbert 2012). The presence of virus-related domains within an LTR/Gypsy element in a *HELPO1.2* copy as well as the occurrence of virus domains in HELPO2 elements suggests that viruses may have participated in the horizontal transfer of these elements from an anonymous ancestor to basidiomycete fungi. The lack of captured genes in the HELPO2 family, along with the above mentioned fact, suggests that fungal helitrons are less likely to capture genes and/or gene fragments than plant and animal helitrons. In fact, none of the intact elements showed any evidence of carrying *P. ostreatus* gene fragments. The captured genes *capD* and *capE* of the *helpo1.3* element also contain animal (the large tegument proteins UL36 of the herpes virus (PHA03247) and plant (Caulimovirus viroplasmin, PF01693) viral sequences. Some researchers have described the occurrence of footprints resulting from EVE integration into host genomes mediated by the retrotransposon enzyme machinery (Feschotte and Gilbert 2012). With the exception of HELPO1.3, the HELPO1 and HELPO2 families contain putative autonomous elements containing three motifs that define the catalytic core as well as the helicase domain. Although fungal RepHel helicases are often described to be intronless (Kapitonov and Jurka 2007), the RNA-seq profiles of the *P. ostreatus* strain N001 revealed the presence of introns in the RepHel genes of the *HELPO1.1* and *HELPO1.2* elements (Fig 5). We did not find any of the previously described domains in the RepHel ORF such as the replication protein A (RPA) found in plant helitrons and occasionally in animals (Jurka 2000), the

zinc fingers present in cnidarian, insect, fish, frog, reptile and mammalian helitrons, or the apurinic (EN) and cysteine protease (CPR) found in cnidarian, fish and frog helitrons. In contrast, a set of conserved domains from viruses, bacteria and eukaryotes never found before in helitrons were present in *P. ostreatus.* The similarity between the RepHel proteins in HELPO1.1 and HELPO1.2 (68.5%) indicates their importance for helitron-specific functions. The similarities between *capA*, *capA2* and *capB* (approximately 45%) suggest that a functional divergence could have occurred, leading to the maintenance (or suppression) of their activities that conferred a possible advantage for the host genome. In this sense, the RT-qPCR experiments showed the highest levels of expression of the *capA* and *capA2* genes in the PC9 and PC15 strains and lower expression levels of *capB*. It should be mentioned that the *capA* gene carried by the *HELPO1.1* elements maps to chromosome VII in a region containing a QTL for earliness and mushroom yield in the dikaryotic strain N001 (R^2 = 32.07).

Phylogenetic reconstruction of Rephel helicases

The helitron helicase-like and Pif1-like helicase domains are present in the putative autonomous elements of every species and are under selective pressure because they are essential for helitron transposition. Thus, these domains retain conserved motifs that can be used to infer the phylogenetic relationships between the helitrons of different organisms. This feature is relevant considering the high variability present within helitron boundaries driven by their ability to capture and reshuffle gene fragments from their hosts. Our phylogenetic analysis revealed a clear polyphyletic origin of these domains, suggesting that horizontal gene transfer played a role in shaping the current distribution of helitrons in extant eukaryotic genomes. Nevertheless, the direction and order of these events cannot be properly assessed given our current sample size. The differential expansion of RepHel helicases in ascomycetes and basidiomycetes, along with the presence of viral domains within helitron boundaries gives strength to the hypothesis of horizontal transfer. In fact, viruses have been proven to be vectors of horizontal transfer of other TEs between eukaryotic hosts sharing viral pathogens (Piskurek and Okada 2007; Routh et al. 2012; Gilbert et al. 2014). An important point to emphasize is that, in addition to plant and animal viruses, bacterial and eukaryotic domains were also found to be integrated into *Pleurotus* helitrons. Previous genomics analyses have shown that HGT could play a more important role in fungal evolution than originally thought (Fitzpatrick 2012). In this regards, previous studies described a bipartite structure similar to that of the *Aspergillus terreus* genome located in a subtelomeric region in *P. ostreatus* suggesting a putative lateral transfer between fungal species (Pérez et al. 2009) . Until now, there was evidence of horizontally transferred helitrons in insect viruses (Thomas et al. 2010), but this is the first report dealing with the presence of viral domains inside helitron transposons. The presence of these

domains in both of the *P. ostreatus* helitron families reinforces their putative role in these transfer events, although reconstructing the phylogenetic history of these elements remains difficult. Based on our data, we hypothesize a putative scenario in which helitrons could have been repeatedly transferred to the fungal kingdom. This horizontal transfer might have been related to previous viral infections of species belonging to the fungal, plant and animal kingdoms with shared ecological niches.

2.5. References

Barbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC, Lal SK (2012) Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics 190:965–975. doi: 10.1534/genetics.111.136176

Bill CA, Summers J (2004) Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. Proc Natl Acad Sci U S A 101:11135–40. doi: 10.1073/pnas.0403925101

Boulé JB, Zakian VA (2006) Roles of Pif1-like helicases in the maintenance of genomic stability. Nucleic Acids Res. 34:4147–4153.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. doi: 10.1093/bioinformatics/btp348

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: The Artemis comparison tool. Bioinformatics 21:3422–3423. doi: 10.1093/bioinformatics/bti553

Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev I V, Stajich JE, Ramírez L (2016) Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet. doi: 10.1371/journal.pgen.1006108

Cultrone A, Domínguez YR, Drevet C, Scazzocchio C, Fernández-Martín R (2007) The tightly regulated promoter of the xanA gene of *Aspergillus nidulans* is included in a helitron. Mol Microbiol 63:1577–1587. doi: 10.1111/j.1365-2958.2007.05609.x

Deng W, Nickle DC, Learn GH, Maust B, Mullins JI (2007) ViroBLAST: A stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics 23:2334–2336. doi: 10.1093/bioinformatics/btm331

Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395:221–36.

Du C, Fefelova N, Caronna J, He L, Dooner HK (2009) The polychromatic Helitron landscape of the maize genome. Proc Natl Acad Sci U S A 106:19916–19921. doi: 10.1073/pnas.0904742106

Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, Blumentritt M, Coutinho PM, Cullen D, De Vries RP, Gathman A, Goodell B, Henrissat B, Ihrmark K, Kauserud H, Kohler A, LaButti K, Lapidus A, Lavin JL, Lee Y-H, Lindquist E, Lilly W, Lucas S, Morin E, Murat C, Oguiza JA, Park J, Pisabarro AG, Riley R, Rosling A, Salamov A, Schmidt O, Schmutz J, Skrede I, Stenlid J, Wiebenga A, Xie X, Kües U, Hibbett DS, Hoffmeister D, Högberg N, Martin F, Grigoriev I V, Watkinson SC (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science (80-) 333:762–765. doi: 10.1126/science.1205411

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi: 10.1093/nar/gkh340

Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13:283–296. doi: 10.1038/nrg3199

Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1:205–20. doi: 10.1093/gbe/evp023

Feschotte C, Wessler SR (2001) Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A 98:8923–4. doi: 10.1073/pnas.171326198

Fitzpatrick DA (2012) Horizontal gene transfer in fungi. FEMS Microbiol. Lett. 329:1–8. doi: 0.1111/j.1574-6968.2011.02465.x

Fu D, Wei L, Xiao M, Hayward A (2013) New insights into helitron transposable elements in the mesopolyploid species *Brassica rapa*. Gene 532:236–245. doi: 10.1016/j.gene.2013.09.033

Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics 100:222–230. doi: 10.1016/j.ygeno.2012.07.004

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–95. doi: 10.1093/oxfordjournals.molbev.a025808

Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, Herniou E a, Cordaux R (2014) Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat Commun 5:3348. doi: 10.1038/ncomms4348

Grigoriev I V, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res 42:D699–D704. doi: 10.1093/nar/gkt1183

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59:307–321. doi: 10.1093/sysbio/syq010

Han MJ, Shen YH, Xu MS, Liang HY, Zhang HH, Zhang Z (2013) Identification and evolution of the silkworm helitrons and their contribution to transcripts. DNA Res 20:471–484. doi: 10.1093/dnares/dst024

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T (2008) PhylomeDB: A database for genome-wide collections of gene phylogenies. Nucleic Acids Res. doi: 10.1093/nar/gkm899

Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11:24. doi: 10.1186/1471-2105-11-24

Jurka J (2010) Helitron elements from moss. *Repbase Reports* 10:961-961

Jurka J (2000) Repbase Update - a database and an electronic journal of repetitive elements. Trends Genet 16:418–420. doi: 10.1016/S0168-9525(00)02093-X

Kapitonov V V, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98:8714–8719. doi: 10.1073/pnas.151269298

Kapitonov V V, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521–529. doi: 10.1016/j.tig.2007.08.004

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–66. doi: 10.1093/nar/gkf436

Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet. doi: 10.1371/journal.pgen.1001191

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858. doi: 10.1038/nprot.2015.053

Koonin E V, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin A V, Sverdlov A V, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7. doi: 10.1186/gb-2004-5-2-r7

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306. doi: 10.1093/bib/bbn017

Labbe J, Murat C, Morin E, Tuskan GA, Le Tacon F, Martin F (2012) Characterization of transposable elements in the ectomycorrhizal fungus Laccaria bicolor. PLoS One 7:e40197. doi: 10.1371/journal.pone.0040197

Lal SK, Hannah LC (2005) Helitrons contribute to the lack of gene colinearity observed in modern maize inbreds. Proc Natl Acad Sci U S A 102:9993–4. doi: 10.1073/pnas.0504713102

Landan G, Graur D (2007) Heads or tails: A simple reliability check for multiple sequence alignments. Mol Biol Evol 24:1380–1383. doi: 10.1093/molbev/msm060

Larraya LM, Perez G, Penas MM, Baars JJP, Mikosch TSP, Pisabarro AG, Ramirez L (1999) Molecular Karyotype of the White Rot Fungus *Pleurotus ostreatus*. Appl Envir Microbiol 65:3413–3417

Larraya LM, Pérez G, Ritter E, Pisabarro AG, Ramírez L (2000) Genetic linkage map of the edible basidiomycete *Pleurotus ostreatus*. Appl Environ Microbiol 66:5290–5300. doi: 10.1128/AEM.66.12.5290-5300.2000

Lassmann T, Sonnhammer ELL (2005) Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6:298. doi: 10.1186/1471-2105-6-298

Li Y, Dooner HK (2009) Excision of Helitron transposons in maize. Genetics 182:399–402. doi: 10.1534/genetics.109.101527

Marchler-Bauer A, Bryant SH (2004) CD-Search: Protein domain annotations on the fly. Nucleic Acids Res. doi: 10.1093/nar/gkh454

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002. doi: 10.1038/ng1615

Pérez G, Pangilinan J, Pisabarro AG, Ramírez L (2009) Telomere organization in the ligninolytic basidiomycete *Pleurotus ostreatus*. Appl Environ Microbiol 75:1427–1436. doi: 10.1128/AEM.01889-08

Piskurek O, Okada N (2007) Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc Natl Acad Sci U S A 104:12046–51. doi: 10.1073/pnas.0700531104

Poulter RTM, Goodwin TJD, Butler MI (2003) Vertebrate helentrons and other novel Helitrons. Gene 313:201–212. doi: 10.1016/S0378-1119(03)00679-6

Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci 104:1895–1900. doi: 10.1073/pnas.0609601104

Ramakers C, Ruijter JM, Lekanne Deprez RH, Moorman AFM (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339:62–66. doi: 10.1016/S0304-3940(02)01423-4

Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, Levasseur A, Lombard V, Morin E, Otillar R, Lindquist EA, Sun H, LaButti KM, Schmutz J, Jabbour D, Luo H, Baker SE, Pisabarro AG, Walton JD, Blanchette RA, Henrissat B, Martin F, Cullen D, Hibbett DS, Grigoriev I V (2014) Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A 111:9923–9928. doi: 10.1073/pnas.1400592111

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26. doi: 10.1038/nbt.1754

Routh A, Domitrovic T, Johnson JE (2012) Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proc Natl Acad Sci U S A 109:1907–1912. doi: 10.1073/pnas.1116168109

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi: 10.1038/msb.2011.75

Thomas J, Schaack S, Pritham EJ (2010) Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol Evol 2:656–664. doi: 10.1093/gbe/evq050

Toleman MA, Bennett PM, Walsh TR (2006) Common regions e.g. orf513 and antibiotic resistance: IS91-like elements evolving complex class 1 integrons. J Antimicrob Chemother 58:1–6. doi: 10.1093/jac/dkl204

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. doi: 10.1093/bioinformatics/btp120

Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34:1692–1699. doi: 10.1093/nar/gkl091

Xiong W, He L, Lai J, Dooner HK, Du C (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci 111:10263–10268. doi: 10.1073/pnas.1410068111

Yang L, Bennetzen JL (2009a) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci U S A 106:19922–19927. doi: 10.1073/pnas.0908008106

Yang L, Bennetzen JL (2009b) Structure-based discovery and description of plant and animal Helitrons. Proc Natl Acad Sci 106:12832–12837. doi: 10.1073/pnas.0905563106

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing R a (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol Biol 7:152. doi: 10.1186/1471-2148-7-152

Chapter III: Transposable elements *versus* the fungal genome: impact on whole-genome architecture and transcriptional profiles

---

This chapter has been published as: Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev I V., Stajich JE, Ramírez L (2016) Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet. doi: 10.1371/journal.pgen.1006108.

3.1. Introduction

Transposable elements (TEs) are mobile genetic units that colonize prokaryotic and eukaryotic genomes and generate intra- and inter-specific variability by promoting a wide range of genomic alterations, some of which are harmful or even lethal to the host. TEs are highly diverse in terms of structure, coding features and transposition mechanisms. According to these characteristics they are classified in classes, orders and superfamilies (Wicker et al. 2007). TEs can be considered selfish elements that parasitize their host genomes, and eukaryotes have developed defense mechanisms for preventing their expansion. Three mechanisms of TE silencing have been described in fungi: i) repeat-induced point mutations (RIP) (Cambareri et al. 1989), ii) transposon methylation (Goll and Bestor 2005; Zemach et al. 2010), and iii) RNA-mediated gene silencing (quelling and meiotic silencing) (Shiu et al. 2001; Fulci and Macino 2007). Repeat-induced point mutations were originally described in *Neurospora crassa* and have been more recently studied in a broad range of filamentous fungi (Horns et al. 2012; Amselem et al. 2015). Transposon DNA methylation has been increasingly studied in the last few years, and recent genome-wide methylation analyses confirm the importance of this epigenetic mechanism in the control of TE proliferation in fungi (Zemach et al. 2010; Montanini et al. 2014; Jeon et al. 2015). Quelling and meiotic silencing occur through the detection of aberrant RNAs, which trigger RNAi pathway genes to silence. Meiotic silencing occurs when chromosomal regions are unpaired during meiosis, such as when a TE is present in one parent but not in the other. Previous studies have shown that meiotic silencing targets unpaired transposable elements (Wang et al. 2015).

Although TEs were originally considered “junk DNA”, we know today that the activity of these elements has strong consequences for genome architecture and that they are key drivers in rapid shifts in eukaryotic genome size (Hawkins et al. 2006; Wessler 2006). Due to their repetitive nature, TEs can promote chromosomal rearrangements through homologous recombination and alternative transposition (Gray 2000). TE activity can also shape genome function in multiple ways. Transposition events can lead to insertional mutations (Bureau and Wessler 1994), which can modify or disrupt gene expression, as well as generate new proteins by exon shuffling and TE domestication (Morgante et al. 2005; Nefedova et al. 2014). In addition, TEs are powerful sources of regulatory sequences (Thornburg et al. 2006) that can be spread across the genome, rewiring pre-established networks or even creating new ones

(Feschotte 2008). Transposable elements are associated with several classes of small RNAs that regulate the expression of multiple genes at the post-transcriptional level (McCue and Slotkin 2012). These reasons, among others, have transformed the originally underestimated importance of TEs into a new, exciting subject of study. This is especially relevant in fungi because international sequencing efforts are rapidly increasing the availability of genome sequences of divergent species with different lifestyles (Floudas et al. 2012; Kohler et al. 2015). Fungal genomes are generally smaller than those of plants and animals, which greatly facilitates their assembly and annotation. However, the accurate annotation and quantification of transposable elements in a genome are not simple tasks, especially in draft assemblies with many scaffolds. Factors such the divergence between TE copies (due to mutations and rearrangements) or the occurrence of nested elements complicate the annotation process and necessitate the use of different algorithms to achieve reliable results (Lerat 2010; Flutre et al. 2011). With the rapid generation of fungal genomes, TE annotation has typically been performed using different strategies, thus limiting the ability to draw robust conclusions about the differences in TE family expansion in different species when copy differences can be ascribed to either methodological differences or biological variation. Recent comprehensive analyses of fungal TEs have described an exceptional variability in the repeat content (Floudas et al. 2012; Hess et al. 2014; Amselem et al. 2015), in which amplification events tend to be more related to the fungal lifestyle than to phylogenetic proximity. LTR-retrotransposons are usually the most abundant mobile elements in fungal genomes, especially those that belong to the Gypsy and Copia superfamilies. In contrast, DNA elements generally constitute a smaller fraction of the fungal repeats, although in some species such as *Fusarium oxysporum*, they have undergone important amplifications in lineage-specific genomic regions (Ma et al. 2010). In this study, we used a multi-approach pipeline for TE annotation in a collection of fungal genomes of varying phylogenetic distances and a detailed analysis of TEs in two strains of *P. ostreatus.* This species is a white rot basidiomycete fungus that grows on tree stumps in its natural environment, and whose life cycle alternates between monokaryotic (haploid) and dikaryotic (dihaploid) mycelial phases. Our results depict a *P. ostreatus* TE landscape dominated by Class I elements that tend to aggregate in non-homologous clusters. These clusters have profound impacts on the genome architecture at intra and inter-specific levels. In addition, we show that TE insertions modulate the global transcriptome of *P. ostreatus* and other fungi.

3.2 Materials and methods

Fungal genomes

Eighteen Ascomycetes and Basidiomycetes species were selected in this study as sample sets of closely related species for genomes comparisons. Publicly available genomic assemblies were downloaded from the Joint Genome Institute's fungal genome portal MycoCosm (Grigoriev et al. 2014) (http://jgi.doe.gov/fungi), the Broad Institute (https://www.broadinstitute.org/) and FungiDB (Stajich et al. 2012). The genome sequences of the *P. ostreatus* monokaryotic strains PC15 v2.0 (Riley et al. 2014) and PC9 v1.0, which were obtained by de-dikaryotization of the dikaryotic strain N001 (Larraya et al. 1999), were used as models for building the pipelines described in this paper.

Identification, classification and annotation of transposable elements (TEs)

De novo identification of repetitive sequences in the genome assemblies was performed by running the RECON (Bao and Eddy 2002) and RepeatScout (Price et al. 2005) programs (integrated into the RepeatModeler pipeline). LTRharvest (Ellinghaus et al. 2008) was used to improve the detection of full length LTR-retrotransposons. LTRharvest results were filtered to avoid false positives as follows: elements were de-duplicated and used as queries for BLASTN searches (cutoff E-value = 10^{-15}) against the genome assembly and for BLASTX (cutoff E-value = 10^{-5}) against the Repbase peptide database (Jurka 2000) . Only sequences longer than 400 bp with more than five copies or yielding a significant hit to a described LTR-retrotransposon were kept for further analysis. The outputs of the above programs were merged and clustered at 80 % similarity using USEARCH (Edgar 2010) to create species-specific (i.e., *P. ostreatus* PC15 and PC9) or genus-specific (i.e., *F. oxisporum* and *F. graminearum*) TE libraries. Each consensus sequences library was classified using BLASTX against the Repbase peptide database, and the final libraries were used as input for RepeatMasker (http://www.repeatmasker.org). Consensus sequences without similarity to any Repbase entry were labeled as 'unknown'. The RepeatMasker output was parsed using the *One_code_to_find_them_all* script (Bailly-Bechet et al. 2014) to reconstruct TE fragments into full-length copies and estimate the fraction of the genome occupied by each TE family.

To identify solo-LTRs, the left terminal repeat of every autonomous copy was extracted, and a BLASTN against each assembly was performed. The flanking sequences of every hit (5,000 bp, cutoff E-value = 10^{-15}) were extracted and screened for retrotransposon internal

sequences. Solo-LTRs were defined as those hits lacking internal retrotransposon sequences at the flanking sites.

Analysis of TE distribution in *P. ostreatus*

To determine whether TEs were non-randomly distributed, the distribution of inter-TE distances was compared (Mann-Whitney-Wilcoxon text) with that of the inter-element distances of a randomly generated subset of 1,196 elements. In addition, TEs and gene model annotations were merged and used as reference for a hypergeometric test to test for the presence of regions enriched in TEs. The analysis was performed using REEF (Coppe et al. 2006) with a Q-value of 0.05 (FDR 5 %), a window width of 100 kb with a shift of 10 kb and a minimum number of 10 features in clusters.

Whole-genome alignment

The *P. ostreatus* PC15 and PC9 genome assemblies were aligned using the Mercator and MAVID pipeline (Dewey 2007), using the fully assembled PC15 genome as a reference. Gene model positions and TE hits of the PC15 strain were used to extract individual alignments and to check the homozygous *vs*. heterozygous nature of the insertions. A locus was considered homozygous if the alignment spanned at least 80 % of the whole locus length, and heterozygous when the PC9 allele was absent.

Estimation of LTR-retrotransposon insertion dates.

Long Terminal Repeats of every intact, full-length element were extracted and aligned. Kimura 2-Parameter distance was obtained using a Python script and transformed to My using the approach described in (Kasuga et al. 2002) and the fungal substitution rate of 1.05×10^{-9} nucleotides per site per year (Dhillon et al. 2014).

Nucleic acid extraction, manipulation and sequencing

Mycelia were harvested, frozen and ground in a sterile mortar in the presence of liquid nitrogen. DNA was extracted using a Fungal DNA Mini Kit (Omega Bio-Tek, Norcross, GA, USA). Sample concentrations were measured using a Qubit® 2.0 Fluorometer (Life Technologies, Madrid, Spain), and purity was measured using a NanoDrop™ 2000 (Thermo-Scientific, Wilmington, DE, USA). PCR reactions were performed according to (Sambrook et al. 1989) using primers designed to match TE flanking sequences (Supplementary Information: Table S1). Total RNA was extracted from 200 mg of deep frozen tissue using Fungal RNA E.Z.N.A Kit (Omega Bio-Tek, Norcross, GA, USA), and its integrity was

estimated by denaturing electrophoresis on 1% (w/v) agarose gels. Nucleic acid concentrations were measured using a Nanodrop™ 2000 (Thermo Scientific, Wilmington, DE, USA), and the purity of the total RNA was estimated by the 260/280 nm absorbance ratio. Messenger RNA was purified using a MicroPoly(A) Purist kit (Ambion, USA). Transcriptome libraries were generated and sequenced by Sistemas Genomicos S.L. (Valencia, Spain) on a SOLiD platform, following the manufacturers' recommendations (Life Technologies, CA, USA). Raw sequencing data was deposited in NCBI under the BioProject accession PRJNA319793.

RNA-seq data analysis

P. ostreatus RNA-seq datasets corresponding to PC15 and PC9 strains (8.4 and 9.7 million reads in PC15 and PC9, respectively) cultured in SMY medium and harvested during the exponential growth phase, were used to analyze the transcription of genes and TEs. The quality of the SOLiD RNA-seq reads was verified using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and they were mapped to their corresponding PC15 v2.0 or PC9 v1.0 assemblies using TopHat (Trapnell et al. 2009), restricting the multihits option to 1. HTseq-count (Anders et al. 2014) was used to determine the number of reads mapping to every feature. SAMtools (Li et al. 2009), BEDTools (Quinlan and Hall 2010) and custom Python scripts were used to manipulate the data, to calculate RPKMs and to obtain genome coverages. Public RNA-seq data from other species were downloaded from the NCBI SRA database and were analyzed using the same pipeline (accessions SRR1257938 *Saccharomyces cerevisiae* S288C (Wu et al. 2010), SRR1284049 *Botrytis cinerea* B05.10 (Blanco-Ulate et al. 2014), SRR1592424 *F. graminearum* (Sikhakolli et al. 2012) and SRR1165053 *Laccaria bicolor* (Tschaplinski et al. 2014)).

For analyzing the expression of TE families, reads were mapped to the extracted transposon sequences using Bowtie (Langmead et al. 2009) and allowing multi-mapping. RSEM software was used to calculate TE expression because its algorithm is especially designed to handle multi-mapped reads (Li and Dewey 2011). Afterwards, the FPKMs of each family were normalized to the number of elements.

Effect of TE insertions on the expression of downstream genes

Gene and TE annotations were intersected to obtain TE-associated genes (genes overlapping with any TE) and non-TE genes (genes not overlapping with any TE). Afterwards, the closest

TE upstream and downstream to each non-TE gene was obtained at a maximum distance of 1 kb. The resulting genes were organized in three groups: i) genes with an upstream TE, ii) genes with a downstream TE and iii) genes with both upstream and downstream TEs. Control groups were obtained by subtracting target genes (three previous scenarios) to all the non-TE genes.

Phylogenetic analysis of the species used in this study

The predicted proteomes of all species were downloaded from the Mycocosm database (http://genome.jgi.doe.gov/programs/fungi/index.jsf). After all-by-all BLASTP, proteins were clustered with MCL (Enright et al. 2002) using an inflation value of 2. Clusters containing single copy genes of each genome were retrieved (allowing two missing taxa per cluster) and proteins were aligned with MAFFT (Katoh et al. 2002). The alignments were concatenated after discarding poorly aligned positions with Gblocks (Talavera and Castresana 2007). Maximum-likelihood phylogeny was constructed using RaxML (Stamatakis 2014) under PROTGAMMAWAGF substitution model and 100 rapid bootstraps.

Phylogenetic reconstruction of Tc1-Mariner transposases

Using the *P. ostreatus* JGI browser we identified the internal transposase gene of a full length element of TIR_1 family. This protein was used as query for BLASTP searches (cutoff = E^{-5}) against NCBI RefSeq protein database (independent searches were carried out against animal, plant and bacterial databases). The best five animal, plant and bacterial hits were retrieved when possible (only one hit was obtained using plant database). The same search was performed in the JGI database to retrieve the best five basidiomycete hits, and the best five non-basidiomycete hits. Proteins were aligned with MUSCLE (Edgar 2004), and the alignments were trimmed using trimAl (Capella-Gutierrez et al. 2009) with the default parameters. An approximate maximum likelihood tree was constructed using FastTree (Price et al. 2009) and edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/). Transposases from *P. patens*, *Wolbachia* and *Rhizopus oryzae* were further analyzed to exclude the possibility of being a result of database contamination: Using TBLASTN against NCBI Whole-genome shotgun contigs or JGI genomic scaffolds, we identified their genomic position and verified that they were assembled in long scaffolds and surrounded by other host genes.

3.3. Results

TE content in *P. ostreatus*

Two monokaryotic strains of the basidiomycete *P. ostreatus* (PC9 and PC15, (Larraya et al. 1999; Riley et al. 2014)) were used as a model to analyze differences in the occurrence and expansion of transposable element families. We identified and classified 80 TE families based on structural features and homology to previously described elements (Table 1). These families accounted for 6.2 and 2.5% of the total genome size in PC15 and PC9 genomes, respectively. In addition, we found 144 repeat-like consensus sequences that could not be reliably classified and occupied 3.6 and 2.3 % of PC15 and PC9 assemblies, respectively. These elements are referred to hereafter as 'unknown' and were not used in downstream analyses. Our integrated pipeline combined *de novo* predictions of LTRharvest (Ellinghaus et al. 2008) and RepeatModeler (http://www.repeatmasker.org), which were run on the two *P. ostreatus* genomes and merged to obtain a final TE library. This library was used then by RepeatMasker (http://www.repeatmasker.org) to detect and mask TE copies in each genome assembly. Our results showed that the merging strategy clearly outperformed the four independent approaches in terms of the number of detected families (Fig 1A). In fact, none of the TE families could be simultaneously detected by all four approaches, and very few were detected by three. In addition, up to 38 families (48 % of the total) were detected by only one of the four methods. The distribution of family sizes showed that 9 of the 80 families accounted for the N50 repeat fraction in PC15 (50 % of the total TE sequences), whereas 15 families accounted for the N50 repeat fraction in PC9 (Fig 1B). The *P. ostreatus* repetitive element landscape was clearly dominated by Class I transposons, which accounted for 93 % of the total TE content in PC15 and 89 % in PC9. LTR-retrotransposons were the most abundant TE order, and were responsible for the main differences in TE content between PC15 and PC9. In fact, the four largest Gypsy families (Gypsy_1, Gypsy_2, Gypsy_3 and Gypsy_4) accounted for 2.2 % of the PC15 genome size, but only 0.3 % in the case of PC9. In addition, these families displayed 80 full-length copies in the former, whereas only fragments and two full-length copies were found in the latter (Table 1).

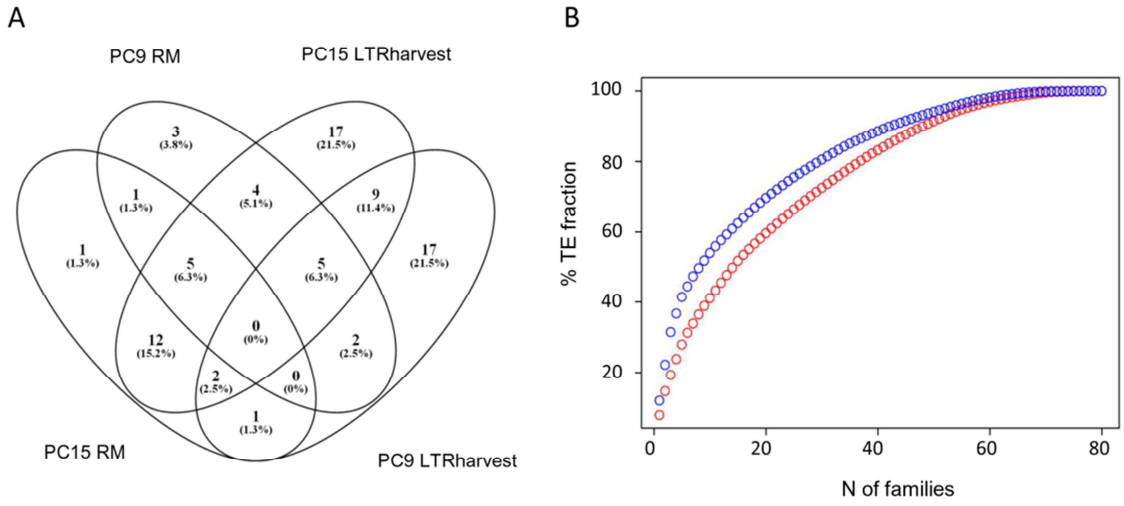

Figure 1. Detection and composition of *P. ostreatus* TE families. Venn diagram showing the number of TE families and their percentage of the total library (in parenthesis) identified in PC15 and PC9 genomes by RepeatModeler (RM) and LTRharvest (A). Cumulative plot showing the number of TE families *vs* total TE fraction (B). PC15 is shown in blue and PC9 in red.

A similar situation occurred with the most prominent Copia families (Copia_1 and Copia_2). Despite the important differences found between PC15 and PC9 in the number of full-length copies and the amount of LTR-retrotransposon masked sequences, the total number of detected TE fragments was closer (1,051 in PC15 *vs* 873 in PC9). The same was true with the amount of solo-LTRs (609 in PC15 *vs* 585 in PC9). Non-LTR retrotransposons (L1 elements) were found in similar abundance in PC9 and PC15, although at lower copy numbers than LTR-retrotransposons. The repertoire of Class II elements found in the genomes was dominated by the previously described Helitron families HELPO1 and HELPO2 (Castanera et al. 2014). In addition, we identified a family of Tc1-mariner transposons (TIR_1) showing putative autonomous elements as well as non-autonomous truncated copies. Autonomous elements of the latter family were present in both genomes, encoding a transposase carrying DDE3 endonuclease (pfam13358) and Tc3 transposase (cl09264) domains. Additionally, TIR_1 elements show terminal inverted repeats of 214 nt and generate a 2bp target site duplication (TA) upon insertion.

Table 1. Summary of detected TE families in *P. ostreatus* strains PC15 and PC9.

Family	Classification	Length (kb)	PC15		PC9	
			Copies *	Kb	Copies *	Kb
Copia_1	LTR-retrotransposon/Copia	4.1	17 (7)	48.2	4 (0)	1.9
Copia_2	LTR-retrotransposon/Copia	5.4	19 (5)	36.1	10 (1)	6.8
Copia_3	LTR-retrotransposon/Copia	6.0	32 (2)	27.9	15 (0)	3.9
Copia_4	LTR-retrotransposon/Copia	5.5	17 (1)	24.2	6 (0)	2.0
Copia_5	LTR-retrotransposon/Copia	6.6	8 (3)	20.6	9 (0)	9.8
Copia_6	LTR-retrotransposon/Copia	5.4	6 (3)	19.3	2 (0)	0.4
Copia_7	LTR-retrotransposon/Copia	5.3	5 (2)	11.1	3 (0)	0.6
Copia_8	LTR-retrotransposon/Copia	5.2	4 (2)	11.0	7 (1)	7.6
Copia_9	LTR-retrotransposon/Copia	5.3	5 (1)	8.8	3 (0)	2.8
Copia_10	LTR-retrotransposon/Copia	5.5	2 (1)	5.8	5 (0)	9.5
Copia_11	LTR-retrotransposon/Copia	5.4	3 (1)	5.7	9 (1)	8.2
Copia_12	LTR-retrotransposon/Copia	1.4	17 (1)	4.3	14 (0)	2.4
Copia_13	LTR-retrotransposon/Copia	5.3	3 (0)	4.0	4 (1)	5.6
Copia_14	LTR-retrotransposon/Copia	5.4	2 (0)	2.9	3 (1)	7.9
Copia_15	LTR-retrotransposon/Copia	5.3	2 (0)	2.0	5 (1)	8.0
Copia_16	LTR-retrotransposon/Copia	5.2	3 (0)	0.3	5 (1)	6.0
Copia_17	LTR-retrotransposon/Copia	1.0	0 (0)	0.0	3 (1)	1.5
DIRS_1	LTR-retrotransposon/DIRS	4.8	22 (2)	19.1	14 (0)	6.6
DIRS_2	LTR-retrotransposon/DIRS	3.7	7 (3)	14.1	11 (4)	21.5
DIRS_3	LTR-retrotransposon/DIRS	1.3	6 (1)	3.7	13 (5)	9.2
DIRS_4	LTR-retrotransposon/DIRS	2.0	0 (0)	0.0	1 (1)	2.0
Gypsy_1	LTR-retrotransposon/Gypsy	6.7	56 (31)	252.4	16 (0)	17.0
Gypsy_2	LTR-retrotransposon/Gypsy	6.7	46 (24)	212.5	12 (1)	4.2
Gypsy_3	LTR-retrotransposon/Gypsy	9.4	54 (18)	192.9	64 (0)	59.8
Gypsy_4	LTR-retrotransposon/Gypsy	11.3	26 (7)	109.9	13 (1)	23.2
Gypsy_5	LTR-retrotransposon/Gypsy	7.0	34 (6)	98.2	34 (3)	70.2
Gypsy_6	LTR-retrotransposon/Gypsy	7.2	29 (1)	63.2	40 (0)	40.4
Gypsy_7	LTR-retrotransposon/Gypsy	7.3	39 (7)	59.5	19 (0)	8.8
Gypsy_8	LTR-retrotransposon/Gypsy	12.5	16 (3)	45.0	13 (0)	9.9
Gypsy_9	LTR-retrotransposon/Gypsy	12.1	41 (1)	39.8	49 (0)	18.0
Gypsy_10	LTR-retrotransposon/Gypsy	8.8	29 (1)	33.7	21 (0)	10.4
Gypsy_11	LTR-retrotransposon/Gypsy	6.9	23 (1)	33.4	14 (0)	2.5
Gypsy_12	LTR-retrotransposon/Gypsy	9.3	6 (3)	33.1	3 (0)	0.4
Gypsy_13	LTR-retrotransposon/Gypsy	10.3	14 (3)	32.1	6 (1)	12.1
Gypsy_14	LTR-retrotransposon/Gypsy	9.4	5 (2)	30.8	1 (1)	9.4
Gypsy_15	LTR-retrotransposon/Gypsy	12.9	16 (1)	28.4	14 (1)	29.2
Gypsy_16	LTR-retrotransposon/Gypsy	9.9	8 (1)	25.5	7 (1)	13.0
Gypsy_17	LTR-retrotransposon/Gypsy	3.4	29 (1)	25.2	21 (0)	4.0
Gypsy_18	LTR-retrotransposon/Gypsy	7.7	22 (2)	24.9	11 (0)	6.3
Gypsy_19	LTR-retrotransposon/Gypsy	11.2	3 (2)	22.7	7 (3)	36.2
Gypsy_20	LTR-retrotransposon/Gypsy	9.2	7 (2)	21.8	5 (0)	10.0
Gypsy_21	LTR-retrotransposon/Gypsy	8.9	5 (2)	21.4	4 (0)	22.4
Gypsy_22	LTR-retrotransposon/Gypsy	9.6	4 (2)	20.9	3 (2)	19.3
Gypsy_23	LTR-retrotransposon/Gypsy	9.6	22 (1)	20.2	17 (0)	7.3
Gypsy_24	LTR-retrotransposon/Gypsy	9.5	4 (1)	19.2	3 (2)	19.4
Gypsy_25	LTR-retrotransposon/Gypsy	7.4	31 (1)	17.0	24 (0)	12.8
Gypsy_26	LTR-retrotransposon/Gypsy	4.2	12 (2)	15.7	16 (0)	2.4
Gypsy_27	LTR-retrotransposon/Gypsy	8.0	19 (1)	14.8	20 (1)	12.2

Gypsy_28	LTR-retrotransposon/Gypsy	9.0	4 (1)	14.7	1 (0)	0.3
Gypsy_29	LTR-retrotransposon/Gypsy	10.0	9 (2)	14.6	10 (0)	14.1
Gypsy_30	LTR-retrotransposon/Gypsy	7.8	10 (0)	14.1	13 (1)	19.2
Gypsy_31	LTR-retrotransposon/Gypsy	5.4	27 (0)	11.7	37 (1)	19.3
Gypsy_32	LTR-retrotransposon/Gypsy	11.5	2 (1)	11.6	2 (0)	0.3
Gypsy_33	LTR-retrotransposon/Gypsy	8.3	13 (1)	11.4	7 (1)	10.6
Gypsy_34	LTR-retrotransposon/Gypsy	8.3	7 (1)	11.1	3 (1)	11.0
Gypsy_35	LTR-retrotransposon/Gypsy	9.2	9 (1)	10.3	11 (1)	15.3
Gypsy_36	LTR-retrotransposon/Gypsy	9.6	2 (1)	9.8	1 (0)	0.3
Gypsy_37	LTR-retrotransposon/Gypsy	9.5	2 (1)	9.7	2 (0)	0.2
Gypsy_38	LTR-retrotransposon/Gypsy	7.9	7 (1)	9.6	4 (1)	11.8
Gypsy_39	LTR-retrotransposon/Gypsy	9.1	3 (1)	9.5	2 (0)	0.3
Gypsy_40	LTR-retrotransposon/Gypsy	10.5	14 (1)	9.3	8 (1)	12.2
Gypsy_41	LTR-retrotransposon/Gypsy	4.9	20 (1)	14.1	17 (0)	3.9
Gypsy_42	LTR-retrotransposon/Gypsy	6.1	11 (1)	10.5	13 (0)	1.7
Gypsy_43	LTR-retrotransposon/Gypsy	4.7	14 (1)	6.7	9 (0)	1.1
Gypsy_44	LTR-retrotransposon/Gypsy	1.0	6 (6)	6.3	5 (5)	5.2
Gypsy_45	LTR-retrotransposon/Gypsy	8.4	6 (0)	4.0	14 (1)	11.8
Gypsy_46	LTR-retrotransposon/Gypsy	5.6	5 (0)	1.4	7 (1)	7.1
Gypsy_47	LTR-retrotransposon/Gypsy	9.1	10 (0)	0.8	11 (1)	13.7
Gypsy_48	LTR-retrotransposon/Gypsy	6.2	2 (0)	0.6	1 (1)	6.2
Gypsy_49	LTR-retrotransposon/Gypsy	5.7	4 (0)	0.5	4 (1)	6.3
Gypsy_50	LTR-retrotransposon/Gypsy	8.9	1 (0)	0.3	1 (1)	8.9
Gypsy_51	LTR-retrotransposon/Gypsy	9.7	2 (0)	0.1	3 (1)	10.1
Gypsy_52	LTR-retrotransposon/Gypsy	2.5	2 (1)	0.1	1 (0)	0.0
Gypsy_53	LTR-retrotransposon/Gypsy	2.8	0 (0)	0.0	1 (1)	2.8
LINE_1	Non-LTRretrotransposon/L1	5.4	14 (4)	30.9	17 (4)	39.6
LINE_2	Non-LTRretrotransposon/L1	2.5	23 (2)	13.7	14 (0)	8.2
LINE_3	Non-LTRretrotransposon/L1	3.8	3 (0)	2.1	6 (1)	6.8
HELPO2	DNAtransposon/Helitron	6.4	15 (5)	44.4	20 (2)	24.0
HELPO1	DNAtransposon/Helitron	7.2	14 (6)	44.9	4 (0)	4.2
TIR_1	DNAtransposon/ Tc1-mariner	1.6	10 (3)	7.3	21 (3)	11.4
TOTAL REPEATS			1051 (204)	2119.2	873 (65)	892.7
Genome percentage (known families)				6.20%		2.50%
Genome percentage (unknown repeats)				3.60%		2.30%

* RepeatMasker reconstructed copies. Full-length copies are shown in parenthesis (>90 % length over family consensus).

Estimation of PC9 TE content from 454 sequencing reads

Our screening of TE sequences in *P. ostreatus* genome assemblies uncovered that some of the most important LTR-retrotransposon families of PC15 were under-represented in PC9 (Table 1). We hypothesized that TE content in PC9 could be underestimated in comparison to PC15 due to its lower assembling quality. In order to know whether this TE families were present in the genome but couldn´t be properly assembled, we analyzed the TE content of PC9 clean 454

sequencing reads (read length of 80 to 626 nt, median length of 364 nt). Datasets of 1.58x and 1.76x genome coverages were randomly sampled from two sequenced libraries, and repeat-masked using our curated TE library to provide an unbiased estimation of TE content. The analysis yielded an average TE content of 4.98%, being the amount of sequence masked by each TE family highly correlated between the two datasets ($R^2 = 0.98$). In addition, the results showed that Gypsy_1, Gypsy_2 and Gypsy_3 LTR-retrotransposon families were the most abundant in PC9 genome, similarly to that found in the fully assembled PC15 strain.

TE distribution across the *P. ostreatus* genome

The density of TEs in *P. ostreatus* was highly variable among the twelve chromosomes and regionally within each chromosome (Fig 2). TEs were not randomly distributed over the genome (Mann-Whitney-Wilcoxon p = 2.2e-16), and overlapped frequently with annotated genes (502 in PC15 and 339 in PC9, hereafter referred as "TE-associated genes"). The results of a hypergeometric test performed on the fully assembled PC15 strain revealed that 58 % of the TEs were arranged in retrotransposon-rich clusters showing poor sequence conservation between the two genomes. A total of 2,108 genes out of 12,330 were present in these repeat-rich regions. Of these genes, 70 were annotated as lignocellulose-degrading enzymes such CAZymes, manganese and versatile peroxidases, although their presence in TE clusters was not over-represented in comparison to the whole genome (Fisher p value = 0.52). At an inter-specific level, the impact of TE insertions was even more striking, as the conservation of these transposon-enriched regions drops dramatically compared with other basidiomycetes (Supplementary Information: Fig S1).

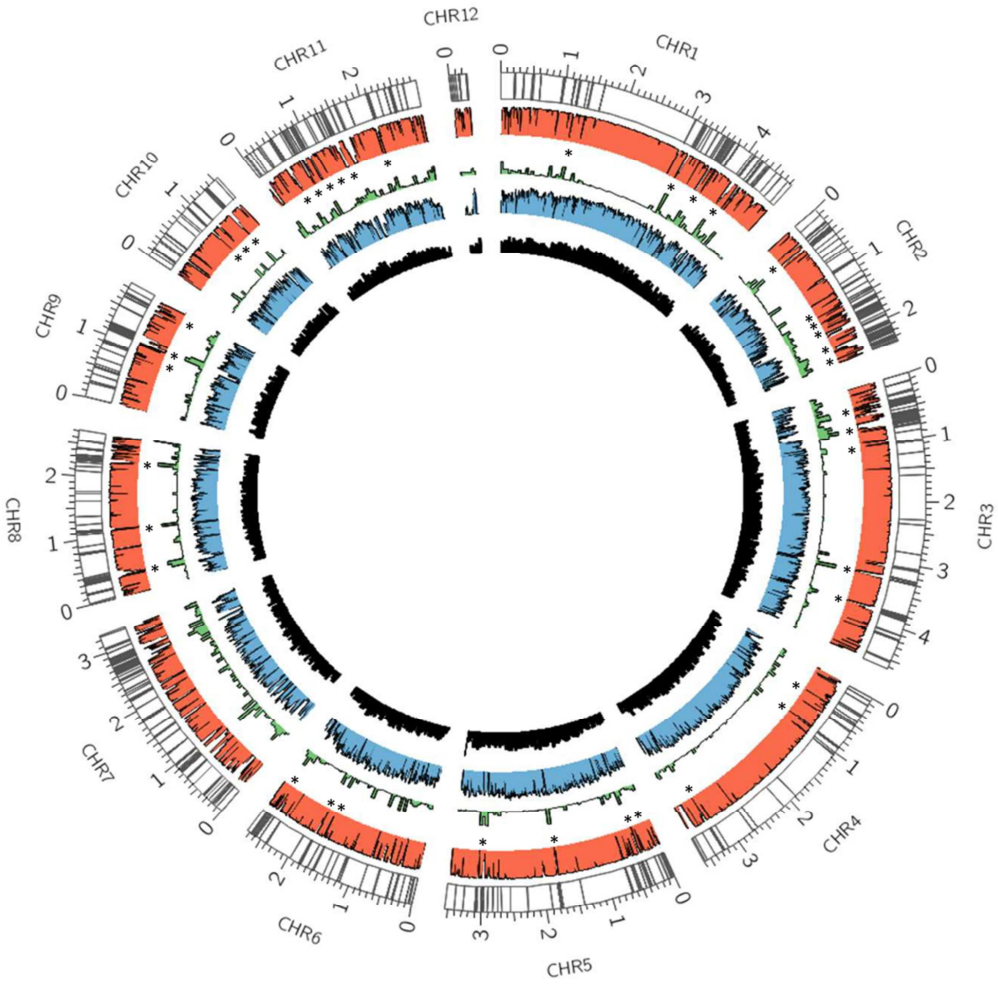

Figure 2. Distribution of transposable elements in the *P. ostreatus* genome and transcriptome context. Each band represents the presence of a transposable element. The PC15 – PC9 genome alignment is shown in red, as a histogram of similarity. Coverage of all repeats (including known and unknown families), transcriptome, and gene densities are shown in green, blue and black histograms. Asterisks indicate regions significantly enriched in TEs ($p < 0.05$).

A whole genome alignment between PC15 and PC9 was performed to detect *in silico* polymorphic TE insertions. The alignment of every TE locus was extracted and parsed to detect the allelic state (genotype) based on the alignability of such regions. We used the same pipeline to analyze the allelic state of 11,630 protein-coding genes. While only 7.7 % of the protein coding genes were heterozygous alleles, up to 50 % of TE insertions were polymorphic. Bioinformatics predictions were validated by PCR in a subset of eight polymorphic insertions (Fig 3).

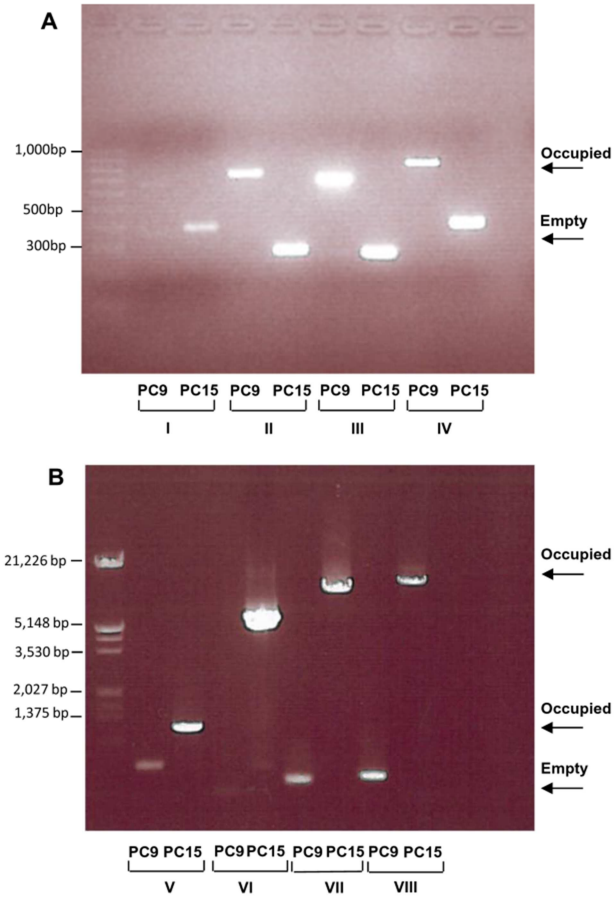

Figure 3. Molecular validation of polymorphic insertions in PC15 and PC9 strains. Primers I to VIII were designed to flank heterozygous TE insertions (present only in one of the two genomes for a given locus) and were used to amplify the target loci in both strains (Supplementary Information: Table S1). Panel (A) shows TE insertions in PC9 strain, and panel (B) shows TE insertions in PC15.

Dynamics of LTR-retrotransposon amplification in *P. ostreatus*

The insertion ages of all intact LTR-retrotransposons (carrying both Long Terminal Repeats, n = 189) were estimated based on the nucleotide divergence of LTRs using the approach described in (SanMiguel et al. 1998) and the fungal substitution rate of 1.05×10^{-9} nucleotides per site per year (Kasuga et al. 2002; Dhillon et al. 2014). Our results showed that 33 % of the LTR-retrotransposon insertions occurred during a recent amplification burst (0 My), and up to 64 % were amplified during the last 5 My (Fig 4).

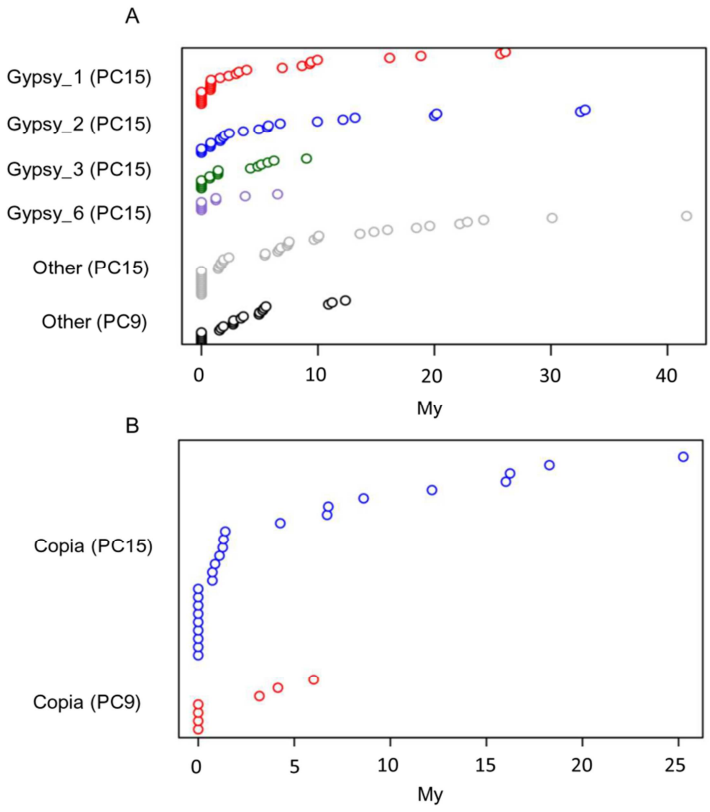

Figure 4. LTR-retrotransposon insertion age in *P. ostreatus*. Estimated insertion dates of Gypsy (A) and Copia (B) elements. Each circle represents one element. Families with more than 5 intact copies have their own category in the Y axes. "Other" represents LTR-retrotransposons belonging to smaller families.

The oldest PC15 LTR-retrotransposon insertion clocked 41 My ago, while the oldest element in PC9 clocked 12 My ago. The phylogenetic reconstruction of the LTR-retrotransposon families revealed that some of the most prominent and recently amplified Gypsy families (Gypsy_1, Gypsy_2, Gypsy_5 and Gypsy_6) were phylogenetically close (Supplementary information: Fig S2).

Transcriptional activity of *P. ostreatus* TEs

We obtained the average expression of every TE family normalized per family size using RNA-seq (Fig 5). Among the main TE groups, LINE was the most abundantly expressed in both strains, followed by Helitrons (especially the HELPO1 family) in PC15 and Gypsy

retrotransposons in PC9. At the family level, 60% were expressed in PC15 and 59% in PC9, while at the copy level only 14 % and 17 % showed transcription, respectively. In addition, 16 out of the 80 families were transcriptionally silent in both strains. Notably, the three strain-specific families in *P. ostreatus* (Copia_17, DIRS_4 and Gypsy_53, present only in PC9) were transcriptionally active.

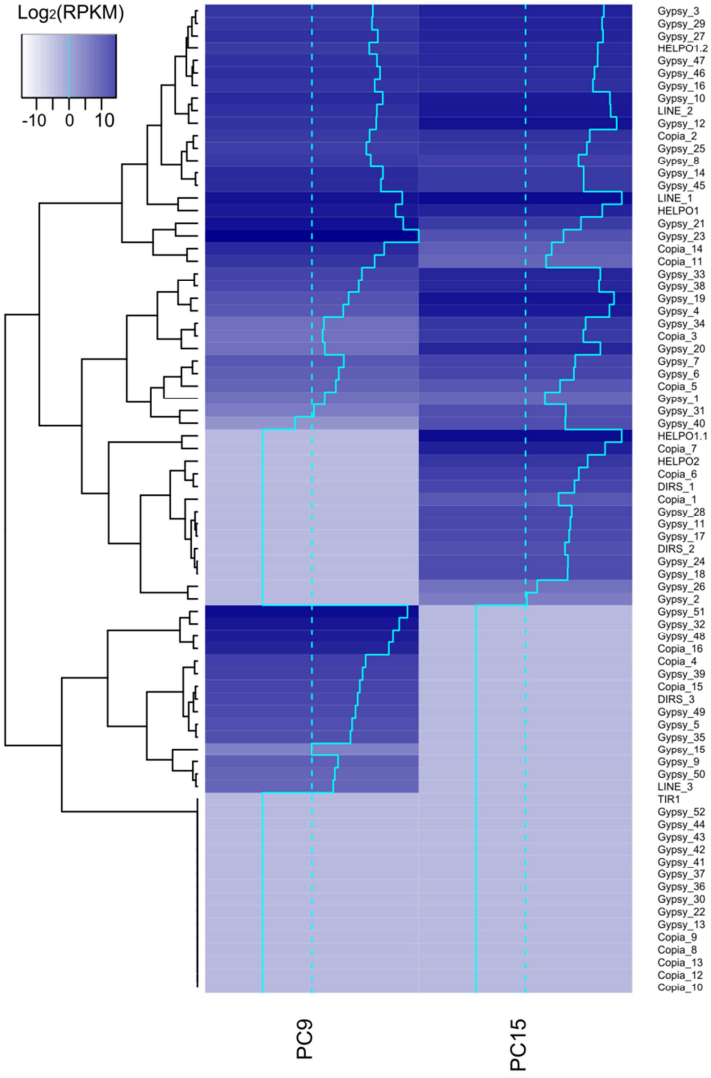

Figure 5. Expression of TE families in *P. ostreatus* PC15 and PC9. Heatmap combined with hierarchical clustering showing the transcription of each TE family in LOG_2(RPKM) normalized per copy number. The blue plain line in the heatmap represents the expression value of each family in the x-axis, and the blue dashed line represents a value of 0 in the x-axis.

Impact of TEs on the *P .ostreatus* functional genome

To investigate the impact of TEs on the functional genome of *P. ostreatus*, we explored the effect of TEs on the expression of the surrounding genes. The closest TE insertion to each gene was identified in the three following scenarios (TE-associated genes were excluded from the analysis): i) a TE was present in a 1kb window upstream of the gene start codon, ii) a TE was present in a 1 kb window downstream of the gene end, and iii) a TE was present in both upstream and downstream regions in a window of 1 kb (gene "captured" between two TEs). This window size was selected based on the small intergenic distance of *P. ostreatus* (1.14 Kb). When we analyzed the gene expression distribution in every scenario, significant differences were uncovered between controls and genes under TE influence (Fig 6A, Fig 6B). In particular, a strong repression was found for genes captured between two TEs (scenario III), while a discontinuous repression was found when the TE was present upstream or downstream of the gene body (scenarios I and II). In the latter case, distribution shapes indicate that approximately half of the genes were repressed and the other half remained unaltered.

To investigate whether this silencing effect could be influenced by the TE distribution along the chromosomes, we split the analysis of the PC15 strain in two additional scenarios: i) the gene under TE influence was located inside a significant TE cluster (Fig 6C) and ii) the gene under TE influence was located outside a significant TE cluster (isolated TE) (Fig 6D). The results showed that the impact of TEs on gene expression was more intense when insertions occurred inside TE clusters. Additionally, significant differences were found between the distribution of gene expression of genes inside clusters that were not under the influence of TEs (control plot, Fig 6C) and that of the genes in the same condition but outside TE clusters (control plot, Fig 6D, $p = 1.22e\text{-}8$).

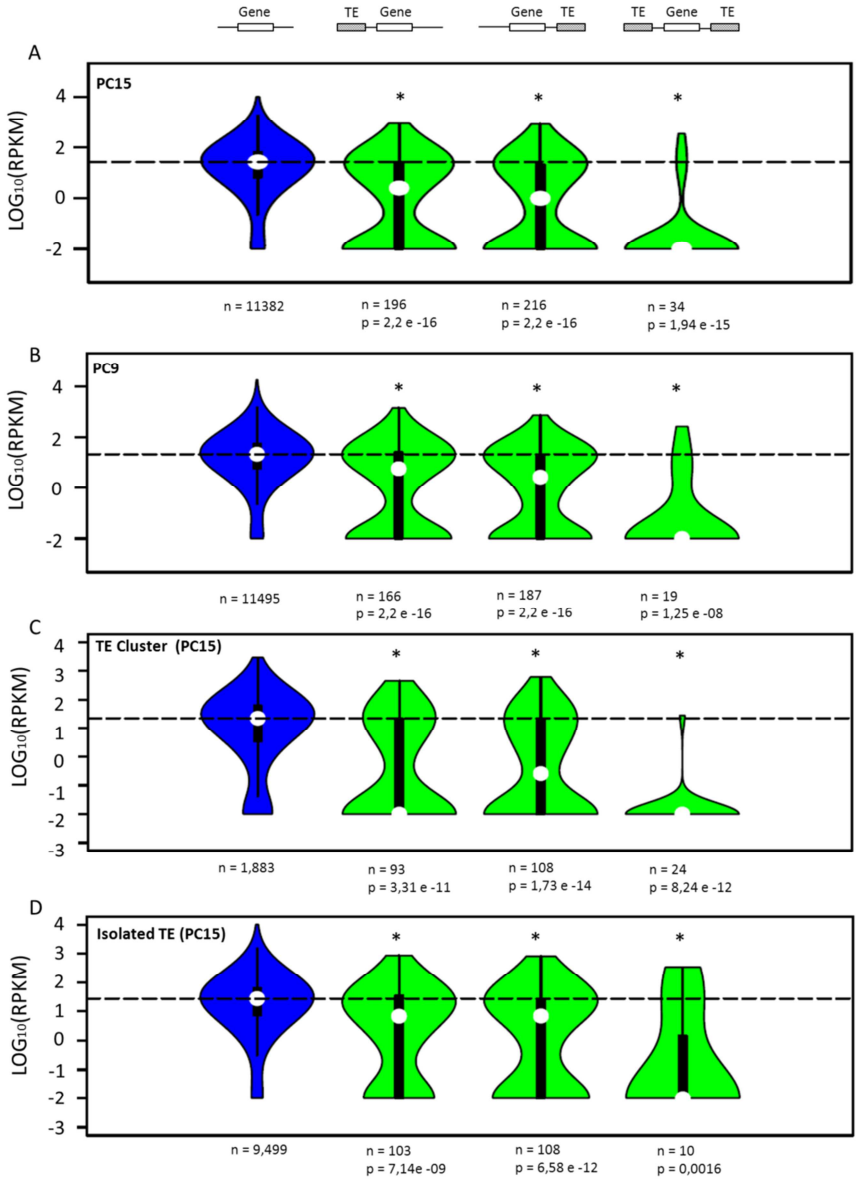

Figure 6. Impact of transposable elements on the expression of neighboring genes in *P. ostreatus*. Green violin plots show the expression of PC15 (A) and PC9 (B) genes carrying a TE insertion in the three studied scenarios. Controls in A and B (blue) show the expression of all non-TE genes that are not represented in the other three scenarios. Chart C shows the expression of PC15 genes inside TE clusters. Control (blue) shows the expression of all non-TE genes localized inside TE clusters that are not represented in the other three scenarios. Chart D shows the expression of genes localized outside TE clusters. Control (blue) shows the expression of all non-TE genes localized outside TE clusters that are not represented in the other three scenarios. For every chart, the dotted line shows the median of the control group. White circles inside violin plots represent the median of each distribution. An asterisk indicates that the gene expression distribution of the test group and the control is different (p

< 0.05, Mann-Whitney-Wilcoxon test). The number of genes belonging to each distribution is shown under the plot (n).

To corroborate the hypothesis of TE-mediated gene repression we studied the transcription of orthologous genes displaying polymorphic insertions (always in a window size of 1 Kb), where a TE was present in PC15 and absent in PC9 and *vice versa*. Table 2 and Table 3 show 21 genes that were inactive under TE influence and active in the orthologous, TE-free allele. Gypsy LTR-retrotransposons were the main TEs involved in the repression with only two exceptions, which involved the Copia_5 (LTR-retrotransposon) and HELPO1 (Helitron) families. The inactivated genes displayed a broad range of functions. Additional orthologous pairs showing strong repression in the allele under TE influence (5 fold) are shown in Table S2 (Supplementary information).

Table 2. Expression of orthologous genes displaying TE insertion in PC15. The first two columns are the protein IDs of the JGI *P. ostreatus* genome database.

PC9 (no TE)	PC15 (TE)	PC9 RPKM	PC15 RPKM	TE family	Interpro description
101709	1048159	73.6	0	Gypsy_2	Unknown
99511	171575	34.6	0	Gypsy_3	Peptidase M
87521	1085356	9.9	0	Helpo1	Unknown
87521	160117	9.9	0	Gypsy_7	Unknown
63834	1109156	2	0	Gypsy_47	Unknown
67552	1033100	1.6	0	Gypsy_1	Unknown
108646	1103939	1.5	0	Gypsy_3	Unknown

Table 3. Expression of orthologous genes displaying TE insertion in PC9. The first two columns are the protein IDs of the JGI *P. ostreatus* genome database.

PC9 (TE)	PC15 (no TE)	PC9 RPKM	PC15 RPKM	TE Family	Interpro description
131667	1102590	0	31.61	Gypsy_3	Unknown
95320	1077306	0	26.88	Helpo1	NAD-dependent epimerase/dehydratase
66978	159492	0	26.77	Gypsy_3	RNA polymerase II, large subunit
68190	33483	0	19.06	Gypsy_31	Unknown
131853	49007	0	12.93	Gypsy_18	Serine/threonine protein kinase
132116	1110152	0	10.36	Gypsy_17	Phospholipase A2
108952	1081099	0	9.72	Gypsy_26	Protein kinase
131565	166826	0	9.72	Gypsy_9	F-box
68399	165925	0	9.42	Gypsy_17	ATP-dependent RNA helicase
91452	160772	0	8.02	Copia_5	Unknown
64875	1109777	0	7.26	Gypsy_6	Cyclin-like
66851	160925	0	2.83	Gypsy_31	Unknown
125628	1102342	0	2.76	Gypsy_41	alpha/beta-Hydrolases
102080	159538	0	1.48	Gypsy_40	Unknown

Differential expansions of transposable elements in fungi

Our pipeline for the identification, classification and annotation of transposable elements was performed in eighteen Ascomycetes and Basidiomycetes genomes (Fig 7). The results demonstrated great variability in TE content at the phylum, genus and species levels (Fig 7, Supplementary information: Table S3). Elements belonging to 20 different TE superfamilies (11 of Class I and 9 of Class II) were identified and classified into the main groups shown in Fig 7. The genome percentage occupied by these TE families showed a positive correlation with genome size (R^2 = 0.38). Within the genera analyzed, *Serpula* showed a surprisingly high TE content in proportion to its genome size, especially due to LTR-retrotransposon expansions in the Gypsy and Copia superfamilies. In fact, when excluding the two *Serpula* genomes from the analysis, the correlation between TE content and genome size in the remaining species was much higher (R^2 = 0.71). The Ascomycete species analyzed had a ratio of Class I / Class II elements ranging from 0.78 to 4.23 and a low content of repetitive sequences, with the exception of the plant pathogen *F. oxysporum.* Interestingly, this species showed a 15-fold enrichment of transposable elements compared with *F. graminearum* as a result of important expansions of Class II elements (Tc1-mariner and hAT families). The variability in the TE content in the analyzed Basidiomycetes ranged from species practically free of TE repeats, such as in the *Pseudozyma* genera (0.02 % of the genome), to species with

almost one third of their genome masked by the TE library, such as *Serpula lacrymans* or *Puccinia graminis*. TE expansions seemed to be constrained in basidiomycete yeasts such *Pseudozyma* or *Mixia* compared to the rest of the basidiomycetes analyzed. LTR-retrotransposons in the Gypsy and Copia superfamilies families were the main elements responsible for differences in TE content, with the Class I / Class II ratio much higher in basidiomycetes than in ascomycetes (9.3 in average). In fact, these two superfamilies were detected in all species analyzed in this study. When we studied the differential TE amplifications at the genus/species level, we found six pairs that displayed similar content (*Botrytis*, *Cryptococcus*, *Phanerochaete*, *Serpula*, *Pleurotus* and *Pseudozyma*) and two pairs (*Fusarium* and *Puccinia*) that showed important differences between counterparts.

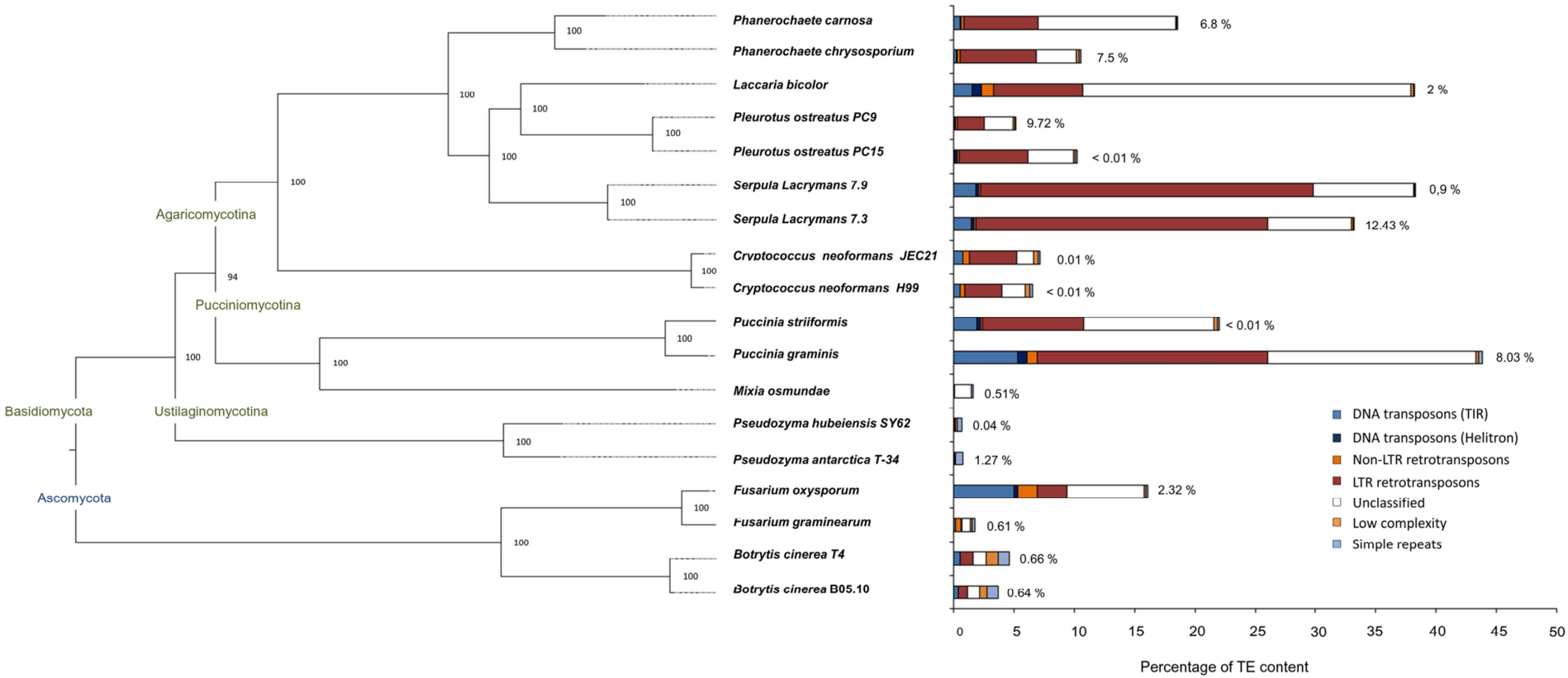

Figure 7. Phylogeny and repeat content of eighteen fungal species. Maximum-likelihood phylogeny inferred with RAxML based on 551 genes and 100 bootstraps. Percentages of assembly gaps are shown near to each bar. Dashed lines are used to align each branch to the tip.

Impact of transposable elements on neighboring gene expression in other fungal models

The effect of TE insertions in nearby genes was analyzed in four additional fungal models: *Laccaria bicolor*, *Fusarium graminearum*, *Botrytis cinerea* B05.10 and *Saccharomyces cerevisiae* S288C. These species were chosen based on the public availability of genomic (full genome sequence) and transcriptomic (RNA-seq) data. In addition, *L. bicolor* and *S. cerevisiae* were chosen based on their opposite methylation patterns (evidence of methylation *vs* absence of methylation, respectively (Zemach et al. 2010). The analysis uncovered two clear profiles. First, *L. bicolor* and *F. graminearum* showed a pattern of TE-mediated repression similar to *P. ostreatus*, in which an important number of genes carrying TE insertions within a 1 kb upstream/downstream window were repressed (Fig 8). Second, *B. cinerea* and *S. cerevisiae* genes under TE influence did not show any alteration in expression, with distributions identical to the control ($p > 0.05$, Fig 8)

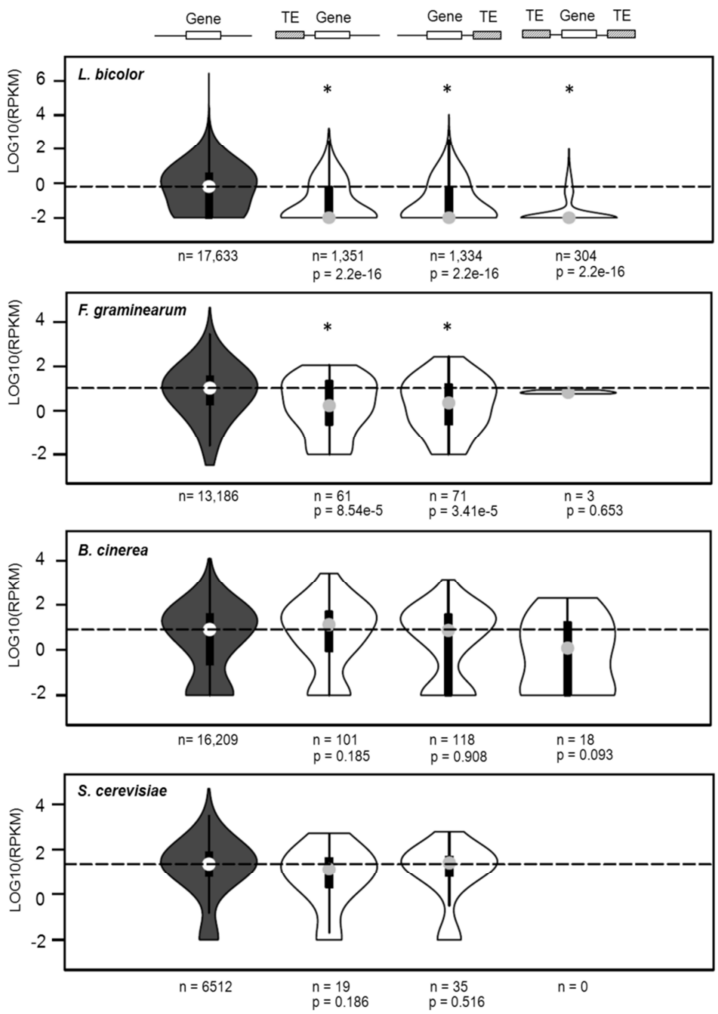

Figure 8. Impact of TE insertions on the expression of the closest gene in four fungal models. *S. cerevisiae* TE annotation was obtained from the SGD database (http://www.yeastgenome.org/). An

asterisk indicates that the gene expression distribution of the test group (white) and the control (grey) is different ($p < 0.05$, Mann-Whitney-Wilcoxon test). The number of genes belonging to each distribution is shown under the plot (n).

Horizontal transfer of Tc1-mariner transposons in eukaryotes

During the process of TE classification using BLASTX against Repbase peptide database we noticed high similarity between the *P. ostreatus* TIR_1 family and the previously described Mariner2_PPa (Jurka 2000), a Tc1-mariner element identified in the moss *Physcomitrella patens* (71 % nucleotide identity over 71 % of the sequence). According to the nucleotide divergence estimated by K2P distance and the fungal nucleotide substitution rate, TIR_1 and Mariner2_PPA diverged 517 My ago, despite mosses and fungi diverged about 1,600 My ago (Heckman et al. 2001). To investigate if horizontal transfers could have played a role in the distribution of fungal and other eukaryotic Tc1-mariners, we reconstructed the phylogeny of their encoded transposases (Fig 9). Our dataset included fungal, animal, plant and bacterial Tc1-mariner transposases, which were obtained based on best BLAST hits against NCBI and JGI reference proteins databases. The topology of the gene tree shows clear incompatibilities with the phylogenetic relationships of the species analyzed, which might be explained by horizontal transfers of Tc1-mariners. Specifically, basidiomycete and animal transposases were placed in a single clade with very high support, separated from ascomycete transposases. Other phylogenetic incongruences were the presence of the moss *Physcomitrella patens* and the mucoral *Rhizopus oryzae* in the basidiomycete clade, as well as the endosymbiont bacteria *Wolbachia* present in the animal clade.

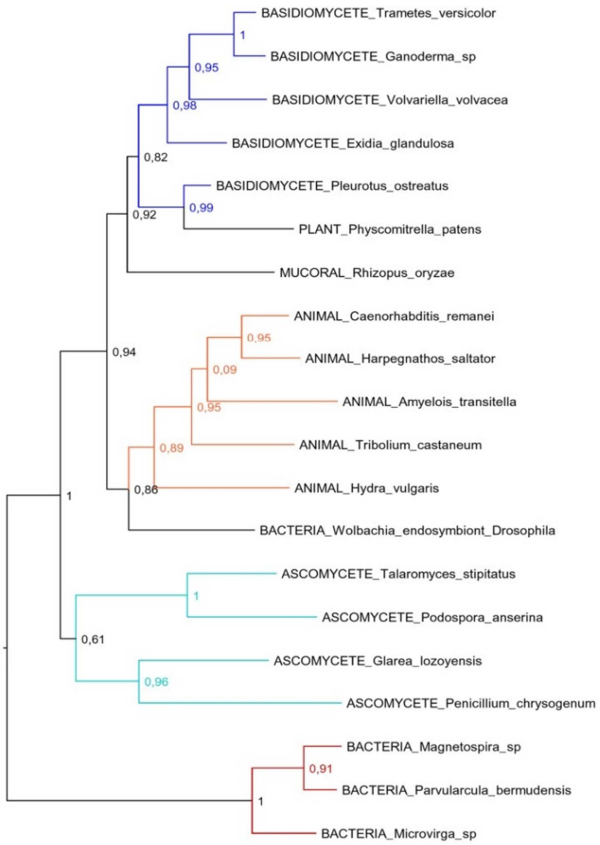

Figure 9. Phylogenetic reconstruction of TIR_1-like Tc1-mariner transposases. Basidiomycete, ascomycete, animal, and bacterial Tc1-mariner transposases are shown in dark blue, light blue, orange and red, respectively. SH indices are included indicating branch support.

3.4. Discussion

TE detection, classification and annotation in *P. ostreatus*

Fungal TE content is highly diverse, even within species that are phylogenetically close (Floudas et al. 2012). However, studies analyzing the intra-specific variability in TE content have been infrequent. According to our results, transposable elements accounted for a small to moderate amount of the genome size in the two *P. ostreatus* strains analyzed (6.2 % in PC15 and 2.5 - 4.9 % in PC9). Although the number of TEs detected varies according to the pipeline used, the TE content in *P. ostreatus* fell within the range reported for most fungal genomes (from 0 to 25%) (Martin et al. 2010; Duplessis et al. 2011; Labbe et al. 2012; Amselem et al. 2015), with the exception of some plant pathogens and ectomycorrhizal species that have undergone massive TE amplifications (Martin et al. 2010; Hess et al. 2014). Despite all TE groups are generally more abundant in PC15 than in PC9, major differences between the strains were observed in LTR-retrotransposons. Most of the LTR-retrotransposon families under-represented in PC9 were actually present in the genome, but could not be assembled into the main scaffolds due to its length and repetitive nature. Assembling transposable elements is technically challenging because identical TE copies require sequencing reads exceeding the TE length to be resolved (McCoy et al. 2014). This is especially relevant in *P. ostreatus*, as we show that most of its LTR-retrotransposons underwent a recent amplification burst, thus sharing high nucleotide similarity. The presence of TE sequences in the unassembled reads is common in plants and animals (Alkan et al. 2011; Hu et al. 2013). In fungi, a recent study performed on several *Amanita* species identified many TEs that could not be found in the assembled regions, especially Gypsy elements (Hess et al. 2014). In addition to the difficulty in assembling TE repeats, their structural complexity, which is caused by internal rearrangements, mutations, nested elements and DNA fragment acquisition events, complicated their identification using generic annotation tools. Our multi-way approach used for TE detection greatly improved the discovery of repeats, as revealed by the number of detected families in our combined TE library (Fig 1A). Using this approach was of particular importance for TE detection in PC9, because families that could not be detected by *de novo* searches in the assembly due to its high gap content could be found in PC15 and thus were present in the TE library.

Transposable element landscape in *P. ostreatus*

P. ostreatus repeat content is enriched in Class I transposons, especially in the Gypsy and Copia superfamilies. LTR-retrotransposons are divided into five superfamilies, but these two are the most abundant in the fungal kingdom (Muszewska et al. 2011; Floudas et al. 2012). The replicative transposition mechanism of autonomous LTR-retrotransposons makes them efficient genome colonizers because the copy number increases with every transposition event. Autonomous LTR-retrotransposons contain *gag* and *pol* genes flanked by long terminal repeats, and they differ from retroviruses in that they do not have infection capacity (Havecker et al. 2004). The difference between the Gypsy and Copia superfamilies lies in the order of the internal protease, integrase, reverse transcriptase and RNAse H domains present in the *pol* gene. We also found retrotransposons of the DIRS superfamily, which contains a *gag*, *pol* and tyrosine recombinase ORFs flanked by terminal repeats. This group of TEs is less abundant than other retrotransposons, and it exhibited patchy distribution in the fungal phylogeny (Muszewska et al. 2013).

One necessary condition for an active TE family is the presence in the genome of autonomous elements encoding the structural features and protein domains necessary for their own transposition. In this sense, the Gypsy architecture seems to be the most successful, as shown by the number of families and number of full-length copies per family. A second condition for TE transposition is that autonomous elements must be transcribed. We showed that although most genomic regions containing TEs are silenced, about 60% of the TE families showed at least one transcriptionally active copy. Interestingly, Class I transposons show high transcriptional levels, which are essential because they are propagated through RNA intermediates that can be translated into proteins necessary for replication or can act as replication templates. In parallel to the successful amplification of LTR-retrotransposons in *P. ostreatus*, the presence of solo-LTRs suggests the occurrence of homologous recombination between LTRs leading to retrotransposons elimination. Class II DNA transposons are less abundant than Class I RNA elements and are represented by the Helitron and Tc1-mariner superfamilies. In a previous work, we reported the presence and structure of the two Helitron families in *P. ostreatus* (Castanera et al. 2014). Helitrons were discovered by bioinformatics approaches in *Arabidopsis thaliana* and *Caenorhabditis elegans* more than a decade ago (Kapitonov and Jurka 2001). Nevertheless, the experimental demonstration of their transposition was not described until very recently (Grabundzija et al. 2016). Their rolling-circle transposition mechanism and their ability to capture and amplify gene fragments make them interesting subjects of study. Helitrons are present in all eukaryotic kingdoms, although they show patchy distribution in some phylogenetic

clades, such as mammals. In plants, they play an important role in genome evolution, introducing functional diversity by creating new genes and isoforms (Barbaglia et al. 2012). In this study, we showed that Helitrons are the most abundant DNA transposons in the *P. ostreatus* genome and are the second superfamily in transcriptional activity. Our results add a piece of evidence to the fact that this superfamily is actively populating the *P. ostreatus* genome. Interestingly, within the 19 described superfamilies of cut and paste DNA transposons, only Tc1-mariner was found in *P. ostreatus*. According to our results, this superfamily would be the most efficient fungal cut and paste transposon, as it is the most represented in the species analyzed. Nevertheless, most of the copies present in *P. ostreatus* are truncated, and the putative autonomous elements encoding transposases are not expressed in the condition tested. Our phylogenetic reconstruction of TIR_1-like Tc1-mariner transposases shows important discordances with organismal phylogenies, suggesting that horizontal transfer has shaped the distribution of these Class II transposons within the eukaryotic kingdom. Specifically, the presence of animal, plant, bacterial, mucoral and basidiomycete transposases in a monophyletic group separated from ascomycetes supports the hypothesis that multiple horizontal transfers occurred after the divergence of basidiomycetes and ascomycetes, event that took place about 1200 My ago (Heckman et al. 2001). It is known that transposable elements are horizontally transferred in eukaryotes at a higher frequency than regular genes (Keeling and Palmer 2008), and this ability allows them to persist in the course of evolution escaping from vertical extinction (Schaack et al. 2010). Our data suggests that horizontal gene transfer has played an important role in the dynamics of eukaryotic Tc1-mariners. Nevertheless, the diversity of TE copies, their repetitive nature and the limitations of the taxonomic sampling make difficult to reconstruct the full evolutionary history of TIR_1-like Tc1-mariner transposases.

Transposable elements in fungi: burden or opportunity?

Most fungal species have streamlined, compact genomes. Owing to international efforts and advances in genome sequencing over the last decade, there is genomic information for nearly 500 fungal species covering most of the fungal phylogenetic diversity, with more being produced (http://1000.fungalgenomes.org). The assembled genome sizes in fungi range from about 2 to 190 Mb, while flow cytometry estimations have uncovered genome sizes of up to 893 Mb in the Pucciniomycotina subphylum (Tavares et al. 2012) (*Gymnosporangium confusum*). The available data demonstrate the impressive variability in fungal genome size, and our results suggest that an important part of this variability could be explained by differential expansions of TEs that seem to be related to the fungal lifestyle. Our results confirm that obligate biotrophs

such *P. graminis* and *P. striiformis* are highly enriched in TEs (Duplessis et al. 2011). By contrast, the (not obligate) biotroph *M. osmundae* is practically free of TEs, similarly to other basidiomycete yeasts such the *P. hubeiensis* and *P. antarctica*. Previous studies have shown that TE-driven expansions have played important roles in the genomes of filamentous plant pathogens (Raffaele and Kamoun 2012). An example of the impact of TEs in host adaptation and pathogen aggressiveness is the *Leptosphaeria genus* (Grandaubert et al. 2014). According to (Raffaele and Kamoun 2012), faster adaptation occurs because genes encoding proteins for host interactions are frequently polymorphic and reside within repeat-rich regions of the genome. Due to the presence of *P. ostreatus* lignin degrading genes within TE clusters, is tempting to hypothesize that TEs could play an important role in the evolution of wood decayers.

Impact of TEs on genome architecture and functionality

Transposable elements are undoubtedly an important source of genetic variation in fungi. As previously found in other fungal species (Labbe et al. 2012), *P. ostreatus* TEs are preferentially arranged in non-homologous genomic regions that display low conservation at both the intraspecific and interspecific levels. These genomic blocks are hotspots for LTR-retrotransposon accumulation, which might target these regions due to specific chromatin structures adopted by pre-existing elements (Garfinkel 2005).

The compatible monokaryotic strains PC9 and PC15 can mate to form a dikaryon, the nuclei of which coexist in the same cell. Thus, the unpaired long blocks of repetitive DNA are unlikely to undergo crossover and are likely inherited as supergenes after meiosis. We show that the transcription of these TE-rich regions tended to be strongly repressed (Fig 2, Fig 6) and we hypothesize that genes with essential functions might eventually be captured and silenced during the formation of these TE clusters, leading to a looseness of fit by the monokaryotic genotypes carrying these genomic regions. Selection against these TE blocks would lead to the loss of these alleles in the course of evolution. On the other hand, the higher plasticity of these repeat regions might create novel opportunities for diversification and adaptation. In addition to the permanent genomic modifications that TEs can promote, we showed that both isolated and clustered TE insertions modulate the expression of surrounding genes. In addition to the disruption-mediated changes originated by TE insertions into promoter regions, there are additional mechanisms by which TEs can alter the expression of surrounding genes. TEs often carry *cis*-regulatory elements that can be spread over the genome (Feschotte 2008). Similarly, LTR-retrotransposons and solo-LTRs contain promoters that can activate the expression of dormant genes (Garfinkel 2005). Additionally, transcripts from full-length TEs can read through into a neighbor gene,

producing spurious transcripts that can be subjected to transcriptional and post-transcriptional control (Slotkin and Martienssen 2007). Finally, TEs can be targeted for heterochromatin formation, thus potentially silencing the transcription of the adjacent gene (Feschotte 2008). Several studies have shown that *Arabidopsis* genes close to TEs had lower expression than the average genome-wide expression (Hollister and Gaut 2009; Wang et al. 2013). Similarly, a recent study showed that the insertion of SINE retrotransposons close to human and mouse gene promoters led to transcriptional silencing mediated by the acquisition of DNA methylation (Estecio et al. 2012). The few studies available on the subject in fungi indicate that methylation targets transposon sequences selectively, leading to TE transcriptional silencing (Zemach et al. 2010; Montanini et al. 2014; Jeon et al. 2015). Although methylation within fungal genes tends to be low, studies in the plant pathogen *Magnaporthe oryzae* showed that genes that were methylated in upstream or downstream regions resulted in lower transcription than un-methylated genes (Jeon et al. 2015). We hypothesize that the transcriptional repression of genes surrounded by TE insertions could be related to the epigenetic status of the given TE. In fact, the discontinuous repression found in *P. ostreatus* genes under TE influence (gene repressed *vs* non-repressed) fits with the putative methylated *vs* non-methylated status of the involved TEs. Although we lack experimental evidence of methylation in PC15 or PC9, the presence in both strains of transcriptionally active homologs of the *Dim-2* DMTase (Supplementary information: Fig S3) responsible for cytosine methylation in fungi (Kouzminova and Selker 2001) suggests that the methylation machinery is active in *P. ostreatus*. In addition to *P. ostreatus*, we used the same transcriptional analysis pipeline in two species with well-known methylation profiles (Zemach et al. 2010): *S. cerevisiae* (methylation-free) and *L. bicolor* (TE regions highly methylated). The expression distribution of *S. cerevisiae* genes under TE influence was identical to the control ($p < 0.05$), while the distribution in *L. bicolor* showed a severe bias towards low expressed genes. Additional analyses performed in other species uncovered that the ascomycetes *F. graminearum* and *B. cinerea* showed different expression patterns for genes under TE influence. Whereas *B. cinerea* genes remained unaltered, the expression in *F. graminearum* genes was lower than the control. Bisulfite sequencing of *Gibberella zeae* (anamorph: *F. graminearum*) showed that this species has low cytosine methylation levels, although it displays related mechanisms of TE silencing, such as RIP and meiotic silencing (Pomraning et al. 2013). Regarding *B. cinerea*, the unique reference found on the subject showed that no or very little methylation occurred in this species, according to HpaII/MspI restriction patterns (Vergara M, Favaron F, Vannacci G 2000). In summary, we show that transposable element dynamics differentially impact fungal genome-wide transcription patterns, likely as a result of the epigenetic machinery evolved to control TE proliferation.

3.5. References

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65. doi: 10.1038/nmeth.1527

Amselem J, Lebrun M-H, Quesneville H (2015) Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics 16:141. doi: 10.1186/s12864-015-1347-1

Anders S, Pyl PT, Huber W (2014) HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169. doi: 10.1093/bioinformatics/btu638

Bailly-Bechet M, Haudry A, Lerat E (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob DNA 5:13. doi: 10.1186/1759-8753-5-13

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–76. doi: 10.1101/gr.88502

Barbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC, Lal SK (2012) Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics 190:965–975. doi: 10.1534/genetics.111.136176

Blanco-Ulate B, Morales-Cruz A, Amrine KC, Labavitch JM, Powell AL, Cantu D (2014) Genome-wide transcriptional profiling of *Botrytis cinerea* genes targeting plant cell walls during infections of different hosts. Front Plant Sci 5:435. doi: 10.3389/fpls.2014.00435

Bureau TE, Wessler SR (1994) Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc Natl Acad Sci U S A 91:1411–1415.

Cambareri EB, Jensen BC, Schabtach E, Selker EU (1989) Repeat-induced G-C to A-T mutations in *Neurospora*. Science (80-) 244:1571–1575. doi: 10.1126/science.2544994

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. doi: 10.1093/bioinformatics/btp348

Castanera R, Pérez G, López L, Sancho R, Santoyo F, Alfaro M, Gabaldón T, Pisabarro AG, Oguiza JA, Ramírez L (2014) Highly expressed captured genes and cross-kingdom domains present in Helitrons create novel diversity in *Pleurotus ostreatus* and other fungi. BMC Genomics 15:1071. doi: 10.1186/1471-2164-15-1071

Coppe A, Danieli GA, Bortoluzzi S (2006) REEF: searching REgionally Enriched Features in genomes. BMC Bioinformatics 7:453. doi: 10.1186/1471-2105-7-453

Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395:221–36.

Dhillon B, Gill N, Hamelin RC, Goodwin SB (2014) The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. BMC Genomics 15:1132. doi: 10.1186/1471-2164-15-1132

Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feaue N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van De Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev I V, Szabo LJ, Martin F (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci U S A 108:9166–9171. doi: 10.1073/pnas.1019315108

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi: 10.1093/nar/gkq625

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. doi: 10.1093/bioinformatics/btq461

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18. doi: 10.1186/1471-2105-9-18

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–84. doi: 10.1093/ nar/30.7.1575

Estecio MR, Gallegos J, Dekmezian M, Lu Y, Liang S, Issa JP (2012) SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. Mol Cancer Res 10:1332–1342. doi: 10.1158/1541-7786.MCR-12-0351

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405. doi: 10.1038/nrg2337

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, De Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, John FS, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev I V, Hibbett DS (2012) The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science (80-) 336:1715–1719. doi: 10.1126/science.1221748

Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526. doi: 10.1371/journal.pone.0016526

Fulci V, Macino G (2007) Quelling: post-transcriptional gene silencing guided by small RNAs in *Neurospora crassa*. Curr Opin Microbiol 10:199–203. doi: 10.1016/j.mib.2007.03.016

Garfinkel DJ (2005) Genome evolution mediated by Ty elements in *Saccharomyces*. Cytogenet Genome Res 110:63–69. doi: 10.1159/000084939

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481–514. doi: 10.1146/annurev.biochem.74.010904

Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, Gogol-Döring A, Kapitonov V, Diem T, Dalda A, Jurka J, Pritham EJ, Dyda F, Izsvák Z, Ivics Z (2016) A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun 7:10716. doi: 10.1038/ncomms10716

Grandaubert J, Lowe RG, Soyer JL, Schoch CL, Van de Wouw AP, Fudal I, Robbertse B, Lapalu N, Links MG, Ollivier B, Linglin J, Barbe V, Mangenot S, Cruaud C, Borhan H, Howlett BJ, Balesdent MH, Rouxel T (2014) Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. BMC Genomics 15:891. doi: 10.1186/1471-2164-15-891

Gray YH (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet 16:461–468. doi:10.1016/S0168-9525(00)02104-1

Grigoriev I V, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I (2014) MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res 42:D699–D704. doi: 10.1093/nar/gkt1183

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225. doi: 10.1186/gb-2004-5-6-225

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261. doi: 10.1101/gr.5282906

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, Hedges SB (2001) Molecular Evidence for the Early Colonization of Land by Fungi and Plants. doi: 10.1126/science.1061457

Hess J, Skrede I, Wolfe BE, LaButti K, Ohm RA, Grigoriev I V, Pringle A (2014) Transposable element dynamics among asymbiotic and ectomycorrhizal *Amanita* fungi. Genome Biol Evol 6:1564–1578. doi: 10.1093/gbe/evu121

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428. doi: 10.1101/gr.091678.109

Horns F, Petit E, Yockteng R, Hood ME (2012) Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi. Genome Biol Evol 4:240–247. doi: 10.1093/gbe/evs005

Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res 23:89–98. doi: 10.1101/gr.141689.112

Jeon J, Choi J, Lee GW, Park SY, Huh A, Dean RA, Lee YH (2015) Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, *Magnaporthe oryzae*. Sci Rep 5:8567. doi: 10.1038/srep08567

Jurka J (2000) Repbase Update - a database and an electronic journal of repetitive elements. Trends Genet 16:418–420. doi: 10.1016/S0168-9525(00)02093-X

Kapitonov V V, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98:8714–8719. doi: 10.1073/pnas.151269298

Kasuga T, White TJ, Taylor JW (2002) Estimation of Nucleotide Substitution Rates in Eurotiomycete Fungi. Mol Biol Evol 19:2318–2324. doi: 10.1093/oxfordjournals.molbev.a004056

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–66. doi: 10.1093/nar/gkf436

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605–18. doi: 10.1038/nrg2386

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canback B, Choi C, Cichocki N, Clum A, Colpaert J, Copeland A, Costa MD, Dore J, Floudas D, Gay G, Girlanda M, Henrissat B, Herrmann S, Hess J, Hogberg N, Johansson T, Khouja HR, LaButti K, Lahrmann U, Levasseur A, Lindquist EA, Lipzen A, Marmeisse R, Martino E, Murat C, Ngan CY, Nehls U, Plett JM, Pringle A, Ohm RA, Perotto S, Peter M, Riley R, Rineau F, Ruytinx J, Salamov A, Shah F, Sun H, Tarkka M, Tritt A, Veneault-Fourrey C, Zuccaro A, Tunlid A, Grigoriev I V, Hibbett DS, Martin F (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet 47:410–415. doi: 10.1038/ng.3223

Kouzminova E, Selker EU (2001) dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*. Embo J 20:4309–4323. doi: 10.1093/emboj/20.15.4309

Labbe J, Murat C, Morin E, Tuskan GA, Le Tacon F, Martin F (2012) Characterization of transposable elements in the ectomycorrhizal fungus *Laccaria* bicolor. PLoS One 7:e40197. doi: 10.1371/journal.pone.0040197

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi: 10.1186/gb-2009-10-3-r25

Larraya LM, Perez G, Penas MM, Baars JJP, Mikosch TSP, Pisabarro AG, Ramirez L (1999) Molecular Karyotype of the White Rot Fungus *Pleurotus ostreatus*. Appl Envir Microbiol 65:3413–3417.

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Hered 104:520–533. doi: 10.1038/hdy.2009

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. doi: 10.1186/1471-2105-12-323

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim WB, Woloshuk C, Xie X, Xu JR, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RA, Chapman S, Coulson R, Coutinho PM, Danchin EG, Diener A, Gale LR, Gardiner DM, Goff S, Hammond-Kosack KE, Hilburn K, Hua-Van A, Jonkers W, Kazan K, Kodira CD, Koehrsen M, Kumar L, Lee YH, Li L, Manners JM, Miranda-Saavedra D, Mukherjee M, Park G, Park J, Park SY, Proctor RH, Regev A, Ruiz-Roldan MC, Sain D, Sakthikumar S, Sykes S, Schwartz DC, Turgeon BG, Wapinski I, Yoder O, Young S, Zeng Q, Zhou S, Galagan J, Cuomo CA, Kistler HC, Rep M (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464:367–373. doi: 10.1038/nature08850

Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J, Anthouard V, Arcioni S, Artiguenave F, Aury J-M, Ballario P, Bolchi A, Brenna A, Brun A, Buee M, Cantarel B, Chevalier G, Couloux A, Da Silva C, Denoeud F, Duplessis S, Ghignone S, Hilselberger B, Iotti M, Marcais B, Mello A, Miranda M, Pacioni G, Quesneville H, Riccioni C, Ruotolo R, Splivallo R, Stocchi V, Tisserant E, Viscomi AR, Zambonelli A, Zampieri E, Henrissat B, Lebrun M-H, Paolocci F, Bonfante P, Ottonello S, Wincker P (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464:1033–1038. doi: 10.1038/nature08867

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS (2014) Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. PLoS One 9:e106689. doi: 10.1371/journal.pone.0106689

McCue AD, Slotkin RK (2012) Transposable element small RNAs as regulators of gene expression. Trends Genet 28:616–623. doi: 10.1016/j.tig.2012.09.001

Montanini B, Chen P-YY, Morselli M, Jaroszewicz A, Lopez D, Martin F, Ottonello S, Pellegrini M (2014) Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. Genome Biol 15:411. doi: 10.1186/s13059-014-0411-5

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002. doi: 10.1038/ng1615

Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR retrotransposons in fungi. PLoS One 6:e29425. doi: 10.1371/journal.pone.0029425

Muszewska A, Steczkiewicz K, Ginalski K (2013) DIRS and Ngaro Retrotransposons in Fungi. PLoS One 8:e76319. doi: 10.1371/journal.pone.0076319

Nefedova LN, Kuzmin I V, Makhnovskii PA, Kim AI (2014) Domesticated retroviral GAG gene in Drosophila: new functions for an old gene. Virology 450-451:196–204. doi: 10.1016/j.virol.2013.12.024

Pomraning KR, Connolly LR, Whalen JP, Smith KM, Freitag M (2013) Repeat-induced Point Mutation, DNA Methylation and Heterochromatin in *Gibberella zeae* (anamorph: *Fusarium graminearum*). In: D.W B (ed) *Fusarium*: Genomics, Molecular and Cellular Biology.

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:I351–I358. doi: 10.1093/bioinformatics/bti1018

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. doi: 10.1093/molbev/msp077

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi: 10.1093/bioinformatics/btq033

Raffaele S, Kamoun S (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 10:417–430. doi: 10.1038/nrmicro2790

Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, Levasseur A, Lombard V, Morin E, Otillar R, Lindquist EA, Sun H, LaButti KM, Schmutz J, Jabbour D, Luo H, Baker SE, Pisabarro AG, Walton JD, Blanchette RA, Henrissat B, Martin F, Cullen D, Hibbett DS, Grigoriev I V (2014) Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A 111:9923–9928. doi: 10.1073/pnas.1400592111

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, Cold Sprin.

SanMiguel P, Gaut BS, Tikhonov a, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45. doi: 10.1038/1695

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–46. doi: 10.1016/j.tree.2010.06.001

Shiu PK, Raju NB, Zickler D, Metzenberg RL (2001) Meiotic silencing by unpaired DNA. Cell 107:905–916. doi: 10.1016/S0092-8674(01)00609-2

Sikhakolli UR, Lopez-Giraldez F, Li N, Common R, Townsend JP, Trail F (2012) Transcriptome analyses during fruiting body formation in *Fusarium graminearum* and *Fusarium verticillioides* reflect species life history and ecology. Fungal Genet Biol 49:663–673. doi: 10.1016/j.fgb.2012.05.009

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285. doi: 10.1038/nrg2072

Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, Harb OS, Kissinger JC, Li W, Nayak V, Pinney DF, Stoeckert Jr. CJ, Roos DS (2012) FungiDB: an integrated functional genomics database for fungi. Nucleic Acids Res 40:D675–81. doi: 10.1093/nar/gkr918

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–3. doi: 10.1093/bioinformatics/btu033

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–77. doi: 10.1080/10635150701472164

Tavares S, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, Abranches R, Silva Mdo C, Voegele RT, Loureiro J, Talhinhas P (2012) Genome size analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci 5:422. doi: 10.3389/fpls.2014.00422

Thornburg BG, Gotea V, MakaÅ‚owski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110. doi: 10.1016/j.gene.2005.09.036

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. 10.1093/bioinformatics/btp120

Tschaplinski TJ, Plett JM, Engle NL, Deveau A, Cushman KC, Martin MZ, Doktycz MJ, Tuskan GA, Brun A, Kohler A, Martin F (2014) *Populus trichocarpa* and *Populus deltoides* Exhibit Different Metabolomic Responses to Colonization by the Symbiotic Fungus *Laccaria bicolor*. Mol Plant-Microbe Interact MPMI 546:546–556. doi: 10.1094/MPMI-09-13-0286-R

Vergara M, Favaron F, Vannacci G PS (2000) Search for DNA methylation in *Cryphonectria parasitica*, *Botrytis cinerea* and *Pyrenophora graminea*. In: 5th Congress of the European Foundation for Plant Pathology. Taormina.

Wang X, Weigel D, Smith LM (2013) Transposon variants and their effects on gene expression in *Arabidopsis*. PLoS Genet 9:e1003255. doi: 10.1371/journal.pgen.1003255

Wang Y, Smith KM, Taylor JW, Freitag M, Stajich JE (2015) Endogenous Small RNA Mediates Meiotic Silencing of a Novel DNA Transposon. G3. doi: 10.1534/g3.115.017921

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A 103:17600–17601. doi: 10.1073/pnas.0607612103

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi: 10.1038/nrg2165

Wu CY, Rolfe PA, Gifford DK, Fink GR (2010) Control of transcription by cell size. PLoS Biol 8:e1000523. doi: 10.1371/journal.pbio.1000523

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science (80-) 328:916–919. doi: 10.1126/science.1186366

Chapter IV: Genome sequencing and annotation of the basidiomycete *Coniophora olivacea*

This chapter has been submitted as: Castanera R, Pérez G, López-Varas L, Haridas S, Amselem J, Grigoriev I V, Pisabarro AG, Ramírez L (2017) Comparative genomics of *Coniophora olivacea* reveals different waves of genome expansion in Boletales

4.1. Introduction

Coniophora olivacea is a basidiomycete fungus belonging to the order Boletales. *C. olivacea* produces brown rot decay on dead wood of conifers (softwood) and, less frequently, on hardwood species. *C. olivacea* also frequently damages wood buildings or construction materials. The genome sequence of its sister species *C. puteana* was made public in 2012 (Floudas et al. 2012) and contributed to the understanding of genomic differences between brown and white rot fungi (Riley et al. 2014). The Boletales order comprises a diverse group of species including saprotrophs and ectomycorrhizal species such as *Suillus sp.* or *Pisolithus sp.* During the last six years, up to ten Boletales genomes have been sequenced and annotated (Eastwood et al. 2011; Floudas et al. 2012; Kohler et al. 2015). Information emerged from these studies and showed important differences in genomic characteristics between the species belonging to this group, whose predicted common ancestor was dated 84 million years ago. Evolution from this boletal ancestor (supposed to be a brown rot saprotroph) lead to the diversification and the appearance of ectomycorrhizae, which shows a particular contraction of the number of plant cell wall-degrading enzymes coding genes (PCWDE) (Kohler et al. 2015; Martin et al. 2016). In addition, Boletales shows important differences in their genome size and gene content. For example, the smallest assembled boletal genome spans 38.2 Mb and has 13,270 annotated genes (*Hydnomerulius pinastri*), but the largest (*Pisolithus tinctorius*) spans 71.0 Mb and has 22,701 genes (Kohler et al. 2015). Previous studies in saprophytic basidiomycetes have shown that species with higher genome sizes tend to have higher content of more transposable elements (Castanera et al. 2016).

Also, it has been described that species associated with plants (pathogenic and symbiotic) have genomes with expanded TE families (Floudas et al. 2012; Hess et al. 2014), although this trend varies between the three basidiomycete phyla (Castanera et al. 2017). In this paper, we describe the genome sequence and annotation of the brown-rot boletal *C. olivacea*, and we compare it with the genomes of its sister species *C. puteana* as well as with that of two other boletales (*Serpula lacryman*s and *Pisolythus tinctorius*) that have substantially larger genome sizes. The results show than *C. olivacea* displays enzymatic machinery characteristic of brown-rot fungi encoded in a compact genome, carrying a small number of repetitive sequences. The comparative analysis with other Boletales shows that both ancient and modern LTR-retrotransposon amplification events have greatly contributed to the genome expansion along the evolution of Boletales.

4.2. Materials and methods

Fungal strains and culture conditions

Coniophora olivacea MUCL 20566 was obtained from the Spanish Type Culture Collection and was cultured in SMY submerged fermentation as previously described (Castanera et al. 2013).

Nucleic acid extraction

Mycelia were harvested, frozen, and ground in a sterile mortar in the presence of liquid nitrogen. High molecular weight DNA was extracted using the phenol-chloroform protocol described previously (Larraya et al. 1999). DNA sample concentrations were measured using a Qubit® 2.0 Fluorometer (Life Technologies, Madrid, Spain), and DNA purity was measured using a NanoDrop™ 2000 (Thermo-Scientific, Wilmington, DE, USA). DNA quality was verified by electrophoresis in 0.7% agarose gels. Total RNA was extracted from 200 mg of deep-frozen tissue using Fungal RNA E.Z.N.A Kit (Omega Bio-Tek, Norcross, GA, USA), and its integrity was verified using the Agilent 2100 Bioanalyzer system (Agilent Technologies, Santa Clara, CA, USA).

Genome sequencing and assembly

The *C. olivacea* MUCL 20566 genome was sequenced using Illumina ANZPP HiSeq-1TB Regular 2x151 bp 0.309 kb. Each Fastq file was QC filtered for artifact contamination and subsequently assembled with Velvet (Zerbino and Birney 2008). The resulting assembly was used to create a long mate-pair library with an insert size of 3000 +/- 300 bp that was then assembled together with the original Illumina library with AllPathsLG (Gnerre et al. 2010). Raw sequences were deposited in SRA NCBI database under accession number SRP086489.

Transcriptome sequencing and assembly

Strand-specific RNASeq libraries were created and quantified by qPCR. Sequencing was performed using an Illumina HiSeq-2500 instrument. Reads were filtered and trimmed to remove artifacts and low quality regions. Each transcriptome was *de novo* assembled using Trinity (Grabherr et al. 2011) and used to assist annotation and assess the completeness of the corresponding genome assembly using alignments of at least 90% identity and 85% coverage.

Whole-genome alignment

The *C. olivacea* and *C. puteana* assemblies were aligned using the Promer tool from the MUMmer 3.0 package (Kurtz et al. 2004). Genome rearrangements were identified in the Promer output with dnadiff tool.

Genome annotation

The annotation of the *C. olivacea* MUCL 20566 assembled genome was performed using the Joint Genome Institute (JGI) annotation pipeline. First, Repeatmasker (http://www.repeatmasker.org/) was used to detect and mask repeats using the fungal Repbase library (Jurka 2000) as well as a Repeatscout-based library (Price et al. 2005) consisting of highly repetitive elements (> 150 hits). *De novo* assembled transcripts were mapped to the genome using BLAT (Kent 2002) and to NCBI non-redundant protein database using BLASTX (Altschul et al. 1990). Gene prediction was performed with a combination of *ab initio* [FGENESH and GeneMark-ES (Salamov and Solovyev 2000; Ter-Hovhannisyan et al. 2008)]), transcript-based [EST_MAP (http://www.softberry.com/), (Combest unpublished)], and protein-based [GeneWise and FGENESH+ (Birney and Durbin 2000; Salamov and Solovyev 2000)]. The best gene model prediction for each locus was chosen using a custom algorithm.

Functional annotation

The predicted proteome was queried against the NCBI nr (Wheeler et al. 2007), KEGG (Ogata et al. 1999), KOG (Tatusov et al. 2003), Swissprot (Bairoch and Boeckmann 1991) and Pfam (Bateman et al. 2002) databases to assign putative gene functions to the (blastp threshold e^{-5}). In addition, tRNAscan-SE (Lowe and Eddy 1997) was run for predicting tRNAs, Infernal (Nawrocki and Eddy 2013) for identifying putative microRNA precursors, TMHMM (Melén et al. 2003) for transmembrane domains, SignalP (Nielsen et al. 1997) for identifying putative secreted proteins, and InterProScan (Quevillon et al. 2005) for protein domains. Carbohydrate-active enzymes (CAZys) were annotated based on BLAST and HMMER (Johnson et al. 2010) searches against sequence libraries and HMM profiles of the CAZy database (Cantarel et al. 2009) functional modules. Protein structure predictions were carried out with Phyre2 (Kelley et al. 2015).

Annotation of transposable elements

Transposable elements (TEs) were identified and annotated in the genome using REPET pipeline (Quesneville et al. 2005; Flutre et al. 2011). Briefly, *de novo* TE detection was carried out with the TEdenovo module, and the elements were classified with PASTEC. The resulting TE library was fed into TEannot pipeline in two consecutive iterations: the first one with the full library, and the second with an improved library consisting on consensus elements carrying at least one full-length copy after manually discarding false positives (i.e., host genes).

Insertion age of LTR-retrotransposons

Full-length LTR-retrotransposons were identified using LTRharvest (Ellinghaus et al. 2008) followed by BLASTX against repbase. Long Terminal Repeats (LTR) were extracted and aligned with MUSCLE (Edgar 2004). Alignments were trimmed using trimal (Capella-Gutierrez et al. 2009)

and used to calculate Kimura's 2P distance. The insertion age was calculated following the approach described in SanMiguel et al. (1998) using the fungal substitution rate of 1.05×10^{-9} nucleotides per site per year (Dhillon et al. 2014).

Identification gene families

All-by-all BLASTP followed by MCL clustering (Enright et al. 2002) was carried out with *C. olivacea* protein models using a threshold value of e^{-5} and an inflation value of 2. We considered gene families those clusters carrying four or more genes.

Phylogenetic analyses

Species phylogeny was constructed as follows: an all-by-all BLASTP followed by MCL clustering was carried out with a dataset containing the proteomes of all the species. The clusters carrying only one protein per species were identified, and the proteins were aligned using MAFFT (Katoh et al. 2002). The alignments were concatenated after discarding poorly aligned positions with Gblocks (Talavera and Castresana 2007). The phylogeny was constructed using RaxML (Stamatakis 2014) with 100 rapid bootstraps under PROTGAMMAWAGF substitution model. Phylogenetic reconstruction of Gypsy reverse-transcriptases was carried out as follows: Reverse transcriptase RV1 domains were extracted from LTR-retrotransposons of the TE consensus library using Exonerate (Slater and Birney 2005) and aligned with MUSCLE. The alignments were trimmed using trimAl with the default parameters, and an approximate maximum likelihood tree was constructed using FastTree (Price et al. 2009).

4.3. Results

C. olivacea assembly and annotation

The nuclear genome of *C. olivacea* was sequenced with 137 X coverage and assembled into 863 scaffolds accounting for 39.07 Mb. The mitochondrial genome was assembled into two contigs accounting for 78.54 kb. The assembly completeness was 99.78% according to the Core Eukaryotic Genes Mapping Approach (CEGMA), and there was only one missing accession (KOG1322, GDP-mannose pyrophosphorylase). In addition, 97.8% of the sequenced ESTs could be mapped to the genome. The *C. olivacea* assembled genome contained more scaffolds than that of its close relative *C. puteana*, but it had a much lower gap content (Table 1). The total repeat content was 2.91% of which 2.15% corresponded to transposable elements, 0.64% to simple repeats, and 0.12% to low complexity regions. We used transcriptomic information, *ab initio* predictions and similarity searches to annotate a total of 14,928 genes—84.5% of them having a strong EST support (the EST spanning more than 75% of the gene length). In addition, 88.3% of the annotated genes had significant hits to NCBI nr database entries and 46.6% to the manually curated Swiss-Prot database (cutoff e^{-05}). The functional characterization of the *C. olivacea* genes revealed that 6,979 (46.8%) of them had significant homology with members of the KOG, KEGG or GO databases, and 7,841 (52.3%) carried Pfam domains. A total of 1,471 genes (9.8%) carried signal peptide, whereas 470 were predicted to be secreted via the more stringent SECRETOOL pipeline (Cortázar et al. 2014).

The multigene phylogeny based on 1,677 conserved single copy genes displayed different classes, orders and families in branches congruent with previous phylogenetic data (Hibbett et al. 2007), with very high support. *C. olivacea* was placed in a branch along with its sequenced sister species *C. puteana* representing the Coniophoraceae family in the order Boletales (Fig 1).

Table 1. Summary of *C. olivacea* genome sequencing and annotation

Feature	*C. olivacea*	*C. puteana*
Genome assembly size (mB)	39.07	42.97
Sequencing coverage depth	137.7x	49.5x
Number of scaffolds	863	210
Scaffold N50 *	80	7
Scaffold L50 (Mbp) **	0.14	2.40
Nº scaffold gaps	127	412
Genome assembly gaps (%)	0.24%	2.57
Assembly completeness (CEGMA)	99.78%	Unknown
Repeat content (%) ***	2.91%	4.68%
GC content (%)	52.82	52.4
Number of genes	14,928	13,761
Gene density (genes/Mb)	382.07	320.26
Predicted secreted proteins	470 (3.1%)	504 (3.7%)

* N50 indicates the number of scaffolds that account for 50% of the total assembled sequence.

** L50 indicates that 50% of the total sequence is assembled in scaffolds larger than this size.

*** Includes TE, simple repeats and low complexity regions

The whole-genome alignment between the two Coniophoraceae species spanned 52.7% of the *C. olivacea* and 48.0% of *C. puteana* genomes. It shows evidence of macrosynteny between the two species (Fig 2A, Supplementary Information: Fig S1) with an average similarity of 78.4% in the aligned regions (2B) and numerous sequence rearrangements such inversions (565) or translocations (2,414). The good conservation between both genomes in protein coding regions was evidenced by the amount of orthologous genes obtained using the reciprocal best hit approach (7,468 genes with more than 70% identity over 50% of the gene sequence) and by the number of *C. olivacea* proteins yielding significant tBLASTN hits against the *C. puteana* genome (13,572 genes, cutoff e-5, Fig 2C). For the remaining 1,352 *C. olivacea*-specific (orphan) genes, only 48 could be functionally annotated based on KOG, KEGG, GO or InterPro databases.

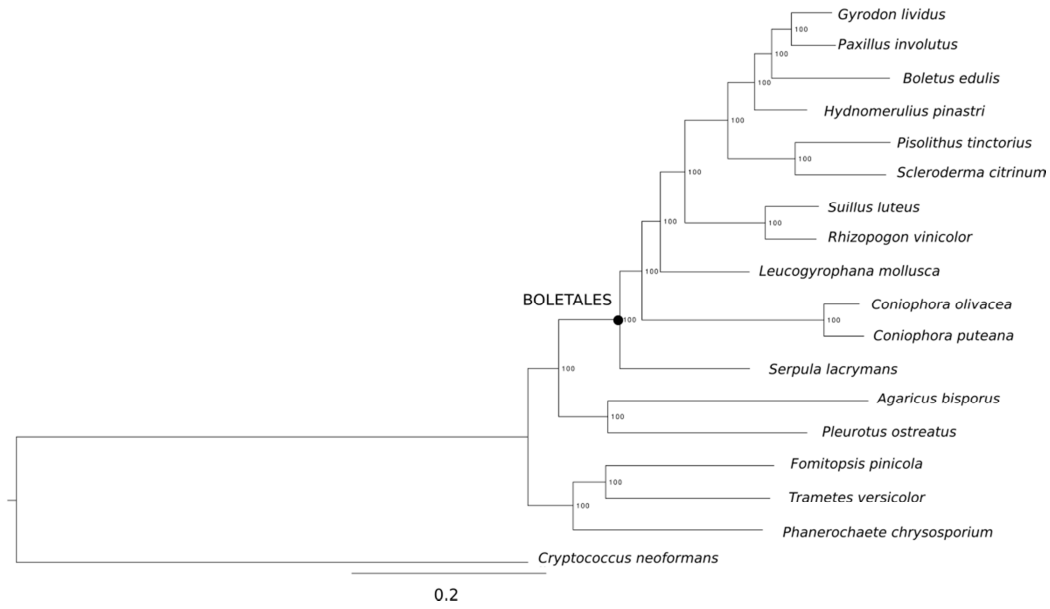

Figure 1. Maximum-likelihood phylogeny of 17 agaricomycetes inferred from 1,677 genes. Branch labels indicate the results of 100 bootstraps.

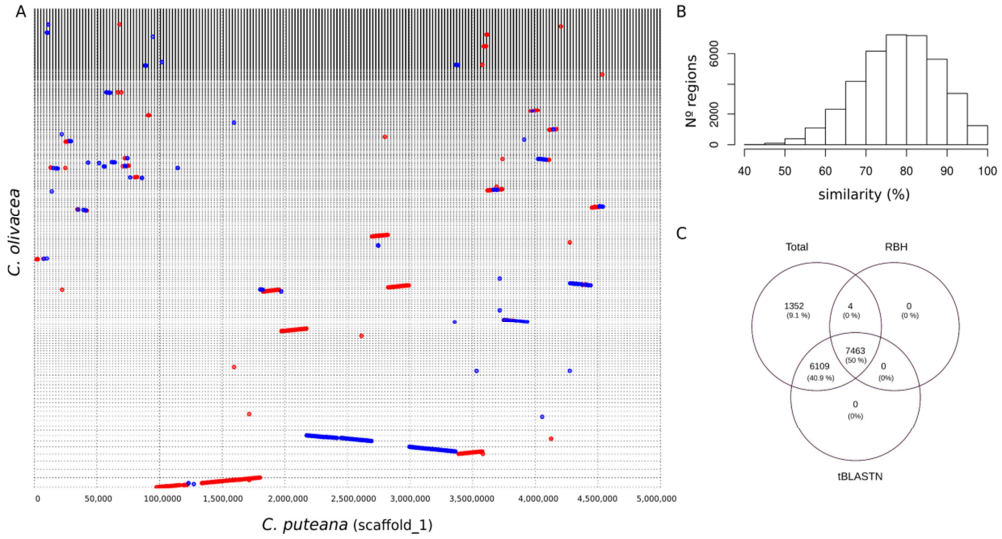

Figure 2. (A) Synteny dot plot showing a fraction of the whole genome alignment between *C. puteana* and *C. olivacea*. Every grid line in the y-axes represents the end of one scaffold and the beginning of the next. Forward matches are displayed in red, while reverse matches are displayed in blue. (B) Histogram of similarity of the 39,506 aligned regions. (C) Venn diagram summarizing the

amount of genes shared by the two genomes based on reciprocal best hit (RBH) and tBLASTN is shown in panel C.

Carbohydrate-active enzymes of *C. olivacea*

The annotated proteome was screened for the presence of carbohydrate-active enzymes (CAZy) using the methods and terminologies described by (Cantarel et al. 2009). A total of 397 proteins were annotated and classified into the different classes and associated modules (Supplementary information: Table S1). The CAZyme profile of *C. olivacea* was very similar to that of *C. puteana* although small differences were found in the glycoside hydrolases. Some families such GH5, GH18 or GH31 were found in lower amounts than in *C. puteana*. Similar to other brown-rot basidiomycetes, *C. olivacea* lacked Class II peroxidases (Auxiliar Activities AA2) and displayed a reduced set of other cellulolytic enzymes such GH6 (1), GH7 (1) and CBM1 (2) and AA9 (6).

Functional characteristics of the predicted secretome

The bioinformatics secretomes of *C. olivacea* and *C. puteana* were predicted using the stringent SECRETOOL pipeline, which considers the presence of signal peptides, cleavage sites, transmembrane domain and the GPI (glycosylphosphatidylinositol) membrane anchor. We used this approach to identify 470 putatively secreted proteins in *C. olivacea* and 504 in *C. puteana*. An enrichment analysis of gene ontology (GO) terms was performed to determine what gene functions were over-represented in the secreted proteins. Thirty GO terms were significantly enriched including 24 corresponding to molecular functions, four to biological processes and two to cellular components. The most enriched molecular function was “feruloyl esterase activity,” which is responsible for plant cell-wall degradation. “Polysaccharide catabolic process” was the most enriched GO term within the biological processes, and “extracellular region” within the cellular components (Table 2).

Table 2. GO terms significantly enriched in the bioinformatics secretome of *C. olivacea*.

Molecular Function	Description	GO/secretome	GO/Genome	p value *
GO:0030600	feruloyl esterase activity	6/470	9/14928	0.000171
GO:0042500	aspartic endopeptidase activity intramembrane cleaving	11/470	20/14928	0.000192
GO:0008843	endochitinase activity	8/470	14/14928	0.000194
GO:0004568	chitinase activity	8/470	14/14928	0.000194
GO:0004650	polygalacturonase activity	11/470	15/14928	0.000354
GO:0004806	triglyceride lipase activity	11/470	29/14928	0.000376
GO:0016160	amylase activity	25/470	40/14928	0.000737
GO:0008933	lytic transglycosylase activity	25/470	40/14928	0.000737
GO:0015927	trehalase activity	25/470	40/14928	0.000737
GO:0015925	galactosidase activity	25/470	40/14928	0.000737
GO:0015924	mannosyl-oligosaccharide mannosidase activity	25/470	40/14928	0.000737
GO:0015929	hexosaminidase activity	25/470	40/14928	0.000737
GO:0015928	fucosidase activity	25/470	40/14928	0.000737
GO:0008810	cellulase activity	9/470	11/14928	0.00089
GO:0015926	glucosidase activity	25/470	41/14928	0.000948
GO:0015923	mannosidase activity	25/470	41/14928	0.000948
GO:0004620	phospholipase activity	9/470	32/14928	0.000968
GO:0004553	hydrolase activity hydrolyzing O-glycosyl compounds	44/470	99/14928	0.00105
GO:0004194	obsolete pepsin A activity	17/470	42/14928	0.00121
GO:0005199	structural constituent of cell wall	16/470	33/14928	0.00129
GO:0030246	carbohydrate binding	9/470	25/14928	0.00143
GO:0004190	aspartic-type endopeptidase activity	20/470	44/14928	0.00193
GO:0004099	chitin deacetylase activity	5/470	9/14928	0.00803
GO:0004185	serine-type carboxypeptidase activity	5/470	12/14928	0.0467
Biological Process				
GO:0000272	polysaccharide catabolic process	5/470	6/14928	0.000414
GO:0006508	proteolysis	43/470	189/14928	0.00128
GO:0005975	carbohydrate metabolic process	65/470	161/14928	0.00176
GO:0006629	lipid metabolic process	10/470	50/14928	0.00674
Cellular Component				
GO:0005576	extracellular region	7/470	15/14928	0.000354
GO:0005618	cell wall	18/470	35/14928	0.00224

* Bonferroni corrected, Fisher p-value

Analysis of putatively secreted multigene families

Using all-by-all BLASTP followed by MCL we clustered the 1,471 proteins carrying signal peptides according to their similarity. As input for similarity clustering we used all proteins carrying signal peptides, to obtain larger protein families. This is because the SECRETOOL pipeline is more stringent. Up to 60% of the 1,471 proteins grouped in clusters were formed by two to 59 genes. When a similar analysis was made using the whole proteome of *C. olivacea*, no differences in the proportions of proteins present in clusters were observed between the two datasets (61% of the 14,928 predicted genes were also found in clusters containing two to 157 members; p=0.6032, Wilcoxon test). For further analysis of the genes found in clusters in the secretome, we focused on 70 clusters (families) formed by four or more members. Using the KOG, KEGG, InterPro and GO databases, we could assign functions to 45 out of the 70 gene families (Table 3). Cytochrome P450,

hydrophobins and aspartic-peptidases were the largest gene families. In addition, 17 CAZys clusters were found including glycoside hydrolases (GH), carbohydrate esterases (CE), carbohydrate-binding modules (CBMs) and redox enzymes classified as auxiliary activities (AA). Nevertheless, we found 25 clusters whose members lacked a functional annotation, and some of them had a high number of genes (clusters 2, 6 and 7 in Table 3). All of these genes belonging to families with unknown function were further analyzed with Phyre2 to predict their protein structure and used for PSI-BLAST (Position-Specific Iterated BLAST) analysis. Using this approach, two gene families were functionally annotated with high confidence (96.3–97.4% confidence for individual protein predictions): one as a copper-dependent lytic polysaccharide monooxygenase (LPMO, also known as AA9; cluster 16), and the other as thaumatin-lyke xylanase inhibitor (*tlxi*, cluster 48).

The Cluster_16 results of putative LPMOs were particularly interesting. These were formed by 10 genes coding for small proteins ranging from 130 to 162 amino acids with three exons (with the exception of protein ID839457 that shows only two). All these genes coded for proteins that have a signal peptide but are not conserved functional domains. Six were confidently annotated as LPMOs by Phyre2, and four were predicted to be secreted by SECRETOOL. In addition, this family of unknown proteins is conserved in all the agaricomycetes shown in Fig 1. Interestingly, four members of this family appear consecutively in a gene cluster located in *C. olivacea* scaffold_124 (scaffold_426:4800-12000).

Table 3. Size and functional annotation of the *C. olivacea* gene families targeted to the secretory pathway according to SignalP and SECRETOOL approaches.

Gene family	SignalP	SECRETOOL	Functional annotation
Cluster_1	59	3	Cytochrome P450
Cluster_2	33	0	Unknown
Cluster_3	32	17	Hydrophobin
Cluster_4	19	11	Aspartic peptidase
Cluster_5	18	12	Carboxylesterase
Cluster_6	17	0	unknown
Cluster_7	15	0	unknown
Cluster_8	14	12	Peptidase G1
Cluster_9	14	9	RlpA-like lipoprotein
Cluster_10	13	0	pheromone mating factor, STE3
Cluster_11	13	0	unknown
Cluster_12	12	3	Peptidase S8/S53
Cluster_13	11	9	unknown
Cluster_14	10	0	CAZY:GH18
Cluster_15	10	9	Cytochrome P450
Cluster_16 *	10	6	**unknown/ lytic polysaccharide monooxygenase (LPMO/ CAZY:AA9)**
Cluster_17	9	5	Aspartic peptidase
Cluster_18	9	5	CAZY:CE4 Carbohydrate Esterase Family 4
Cluster_19	9	0	CAZY:GH16

Cluster_20	9	2	Peptidase S10
Cluster_21	9	5	Sugar transporter
Cluster_22	9	4	unknown/putative lipoprotein
Cluster_23	8	6	Fungal lipase
Cluster_24	8	0	Isoprenylcysteine carboxyl methyltransferase
Cluster_25	8	0	Monooxygenase, FAD-binding
Cluster_26	7	7	Ser-Thr-rich glycosyl-phosphatidyl-inositol-anchored membrane family
Cluster_27	7	0	unknown
Cluster_28	7	1	unknown
Cluster_29	6	5	CAZY:GH128
Cluster_30	6	0	CAZY:GH28
Cluster_31	6	3	CAZY:GH3
Cluster_32	6	2	Peptidase M28
Cluster_33	6	6	Thaumatin
Cluster_34	6	2	unknown
Cluster_35	6	6	unknown
Cluster_36	5	1	Aspartic peptidase
Cluster_37	5	2	CAZY:AA1_1
Cluster_38	5	4	CAZY:AA5_1
Cluster_39	5	5	CAZY:AA9
Cluster_40	5	1	CAZY:CBM5
Cluster_41	5	4	CAZY:GH12
Cluster_42	5	5	CAZY:GH30_3
Cluster_43	5	0	CAZY:GH47
Cluster_44	5	2	CAZY:GH71
Cluster_45	5	0	Monooxygenase
Cluster_46	5	2	unknown
Cluster_47	5	0	unknown
Cluster_48 *	5	5	**Unknown / xylanase inhibitor tl-xi**
Cluster_49	5	4	unknown
Cluster_50	5	4	unknown
Cluster_51	5	4	unknown
Cluster_52	5	0	unknown
Cluster_53	5	4	unknown
Cluster_54	5	0	unknown
Cluster_55	5	0	unknown
Cluster_56	4	0	CAZY:GH18, CAZY:CBM5
Cluster_57	4	0	CAZY:GH31
Cluster_58	4	3	CAZY:GH55
Cluster_59	4	4	Flavin monooxygenase-like
Cluster_60	4	3	GOLD
Cluster_61	4	2	Histidine phosphatase superfamily, clade-2
Cluster_62	4	3	Lysophospholipase
Cluster_63	4	1	Peptidase S28
Cluster_64	4	0	Proteolipid membrane potential modulator
Cluster_65	4	3	RlpA-like, ceratoplatanin
Cluster_66	4	1	Thioredoxin-like fold
Cluster_67	4	3	unknown
Cluster_68	4	3	unknown
Cluster_69	4	3	unknown
Cluster_70	4	0	unknown

Impact of TE content on genome size of species in order Boletales

To study the role that TEs have played in the evolution of the Boletales genomes, we annotated and quantified the TE content in four species showing important differences in genome size: *C. olivacea* (39.1 Mb), *C. puteana* (42.9 Mb) (Floudas et al. 2012), *Serpula lacrymans* (47.0 Mb) (Eastwood et al. 2011) and *Pisolithus tinctorius* (71.0 Mb) (Kohler et al. 2015). The TEs were *de novo* identified and annotated using the REPET pipeline. The results yielded major differences in TE content between the four species with *C. olivacea* and *C. puteana* having very low TE content (2.15% and 3.95% of their corresponding genome sizes), and *S. lacrymans* and *P. tinctorius* having up to 29.45% and 41.17% of their genomes occupied by TEs, respectively (Fig 3, Table 4). In addition to higher TE content, species with larger genome size showed higher TE diversity as reflected by the higher number of TE families.

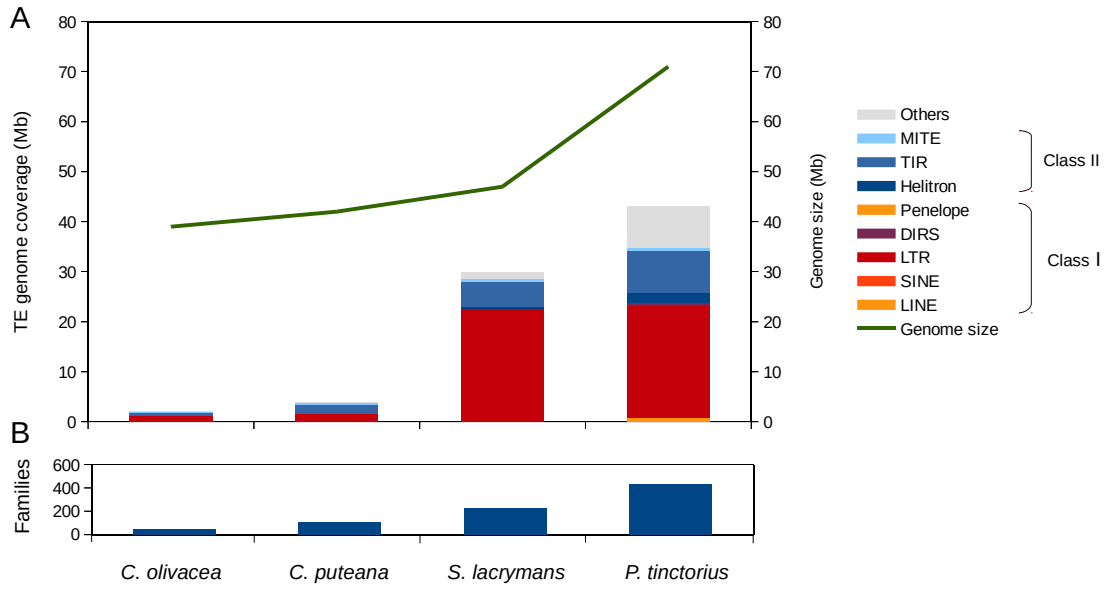

Figure 3. TE content and genome size in four Boletales species. TE content is shown as a histogram, and genome size as a green line in panel A. Panel B shows a histogram representing the number of TE families found in each species.

The TEs belong to six out of the nine TE orders described by Wicker et al. (2007): LTR, DIRS, PLE, LINE (Class I); and TIR and Helitrons (Class II). Two of the orders (LTR and TIRS, TEs containing long terminal repeats or terminal inverted repeats, respectively) were present in four species. Class I TEs were primarily responsible for the observed genome size differences—especially the elements belonging to LTR in the Gypsy superfamily, which accounted for more than 15% of genome size in *S. lacrymans* and *P. tinctorius*, but less than 1% in *C. olivacea* and *C.*

puteana. Of all the LTR/Gypsy families detected by TEdenovo, we observed that those elements belonging to the *Chromoviridae* group (carrying a Chromatin organization domain, PF00385, in the N-terminal region after the integrase, Fig 4) were the most abundant LTR-retrolements in these four species ranging from 44% to 83% of the total Gypsy coverage. The LTR/Copia elements were also particularly abundant in the two *Coniophora* genomes (accounting for 2 – 6 % of the total genome size). Remarkably, non-coding LTR-retrotransposons such as TRIM (terminal-repeat retrotransposons in miniature) and LARD (large retrotransposon derivatives) were also found in three out of the four genomes, but in lower amounts (<1% of the genome, Table 4).

LINE (long interspersed nuclear elements), SINE (small interspersed nuclear elements), DIRS (*Dictyostelium* intermediate repeat sequence) and PLE (penelope-like elements) elements were also found in low copy numbers, but none of these were present in the four species. Regarding Class II orders, TIR was the most important in terms of abundance and copy number with elements encoding DDE transposases present in the four species. The second most important were MITEs (miniature inverted–repeat transposable elements) and other non-coding elements carrying structural features (classified as TIR/unknown in Table 1). Rolling-circle helitrons were only found in *S. lacrymans* and *P. tinctorius*, while Mavericks were present only in this latter one.

Table 4. Summary of TE content in four Boletales genomes

Classification	*C. olivacea* (43 families)			*C. puteana* (108 families)			*S. lacrymans* (230 families)			*P. tinctorius* (432 families)		
	Copies	**Full_copies**	**Coverage (%)**	**Copies**	**Full_copies**	**Coverage (%)**	**Copies**	**Full_copies**	**Coverage (%)**	**Copies**	**Full_copies**	**Coverage (%)**
Class I												
LINE	30	4	0.03	0	0	0.00	0	0	0.00	317	41	0.80
LINE (unknown)	29	5	0.02	11	3	0.01	0	0	0.00	14	1	0.01
SINE	0	0	0.00	6	2	0.00	0	0	0.00	9	1	0.00
LTR/Copia	36	7	0.09	441	27	0.83	3773	86	6.04	1617	101	2.43
LTR/Gypsy	394	13	0.93	299	28	0.54	6949	268	16.27	8434	575	19.28
LTR/LARD	0	0	0.00	60	8	0.08	0	0	0.00	361	2	0.53
LTR/TRIM	15	4	0.02	136	4	0.08	0	0	0.00	576	93	0.20
DIRS	0	0	0.00	0	0	0.00	0	0	0.00	361	36	0.58
Penelope	0	0	0.00	0	0	0.00	69	11	0.15	0	0	0.00
Class II												
Helitron	0	0	0.00	0	0	0.00	260	25	0.43	1386	38	2.01
TIR/DDE	361	28	0.52	362	38	0.68	2148	166	3.04	3366	255	4.25
TIR (unknown)	143	34	0.18	720	115	1.10	736	67	1.55	1115	40	1.85
MITE	410	85	0.30	702	264	0.56	539	98	0.62	1102	227	0.59
Maverick (putative)	0	0	0.00	0	0	0.00	0	0	0.00	56	3	0.21
Unknown	67	4	0.07	99	9	0.06	1138	167	1.34	8611	708	8.44
TOTAL	1485	184	2.15	2836	498	3.94	15612	888	29.45	27325	2121	41.17

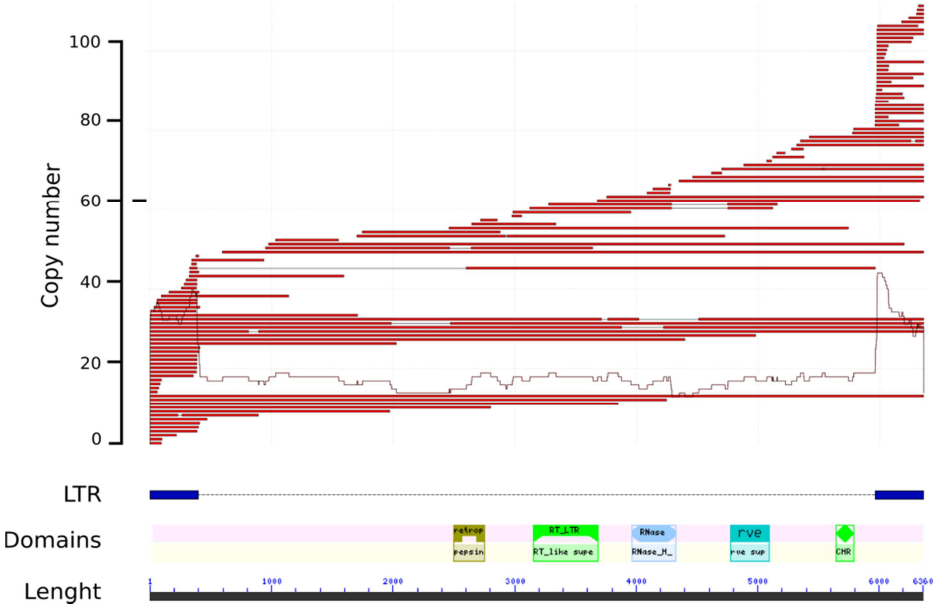

Figure 4. Abundance and structure of a *Chromoviridae* LTR-retrotransposon family of *C. olivacea.* The upper panel shows the mapping of the annotated elements to the family consensus shown in the lower panel. The lower panel shows a scheme of the structural and functional domains of this family: long terminal repeats (LTRs) are represented as blue rectangles; the internal domains shown are (from left to right): aspartate protease, reverse transcriptase, RNase, integrase, chromatin organization modifier.

Phylogenetic reconstruction of the LTR reverse-transcriptases

To understand the phylogenetic relationship between the LTR-retrotransposons in the four analyzed genomes, we inferred a maximum likelihood phylogeny of the LTR reverse-transcriptases of the Gypsy consensus sequences (Fig 5). Three main clades were obtained (A, B and C; Fig 4). Clades A and B were formed, almost exclusively, by elements found in the *P. tinctorius* genome. Moreover,

while clade A appeared to be formed by several distantly related families, the profile of clade B suggests that these families underwent a more recent diversification. All LTR families found in the other three species were analyzed and grouped in clade C along with the remaining families of *P. tinctorius*. Interestingly, clade C contained recently diversified LTR families *S. lacryans* and *P. tinctorium*, which share recent common ancestors with *C. olivacea* and *C. puteana*.

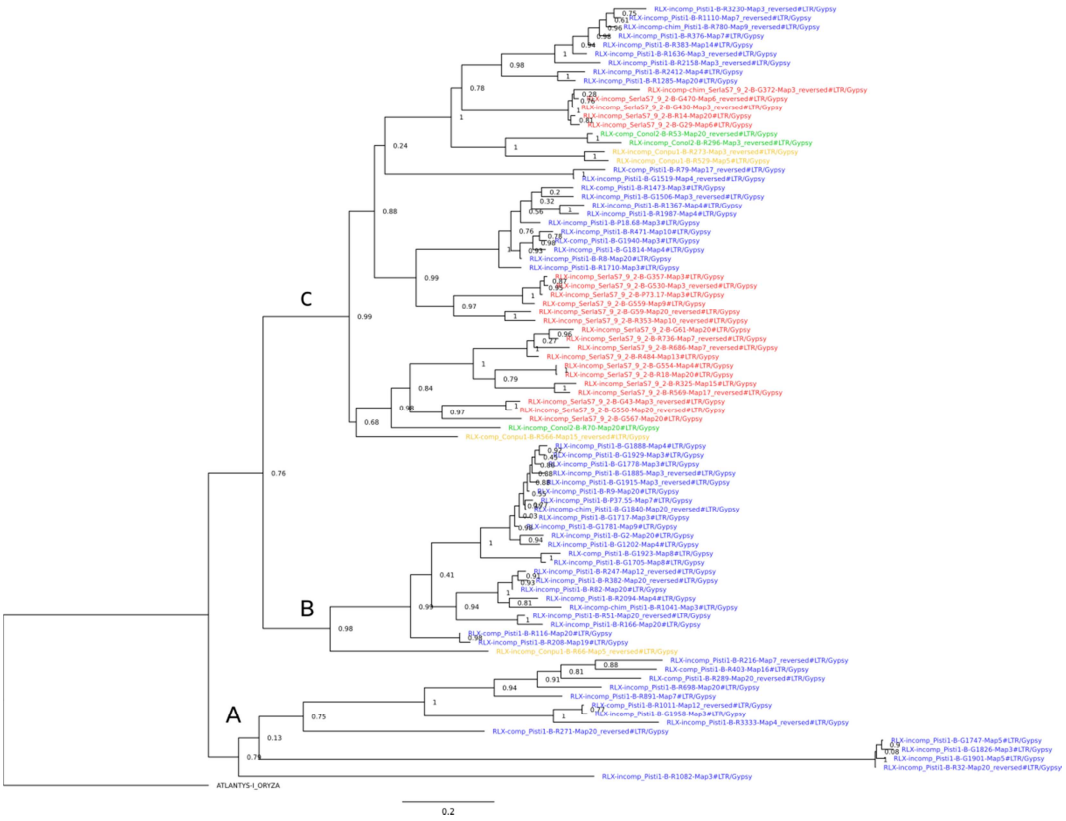

Figure 5. Maximum likelihood phylogeny of the Gypsy reverse-transcriptases found in the *C. olivacea* (green), *C. puteana* (yellow), *S. lacrymans* (red) and *P. tinctorius* (blue) genomes. SH (Shimodaira-Hasegawa) local support values are shown in branches. The reverse-transcriptase from *Oryza sativa* ATLANTIS-I family consensus (Repbase) was used as an out group.

Age of the LTR-retrotransposon amplification bursts in the Boletales

Intact LTR-retrotransposons carrying conserved domains (putative autonomous elements) were subjected to further study to investigate their amplification dynamics over the course of evolution.

Based on the nucleotide divergence between the two LTRs, we estimated the time of insertion of each element using a substitution rate of 1.05×10^{-9} nucleotide substitutions per site per year. The number of full-length, putative autonomous LTR-retrotransposons varied greatly in the four species ranging from 26 elements in *C. olivacea* to 944 in *P. tinctorius.* The LTR profiles of *C. olivacea, C. puteana* and *S. lacrymans* showed recent peaks of amplification with insertion dates at 0-5 million years (MY). In fact, in *C. puteana* and *S. lacrymans*, 32% and 11% of the elements were amplified between 0 and 1 MY ago, respectively. In contrast, the profile of *P. tinctorium* points to a much older amplification burst showing a maximum peak at 30 MY ago and few recent retrotransposition events (Fig. 6).

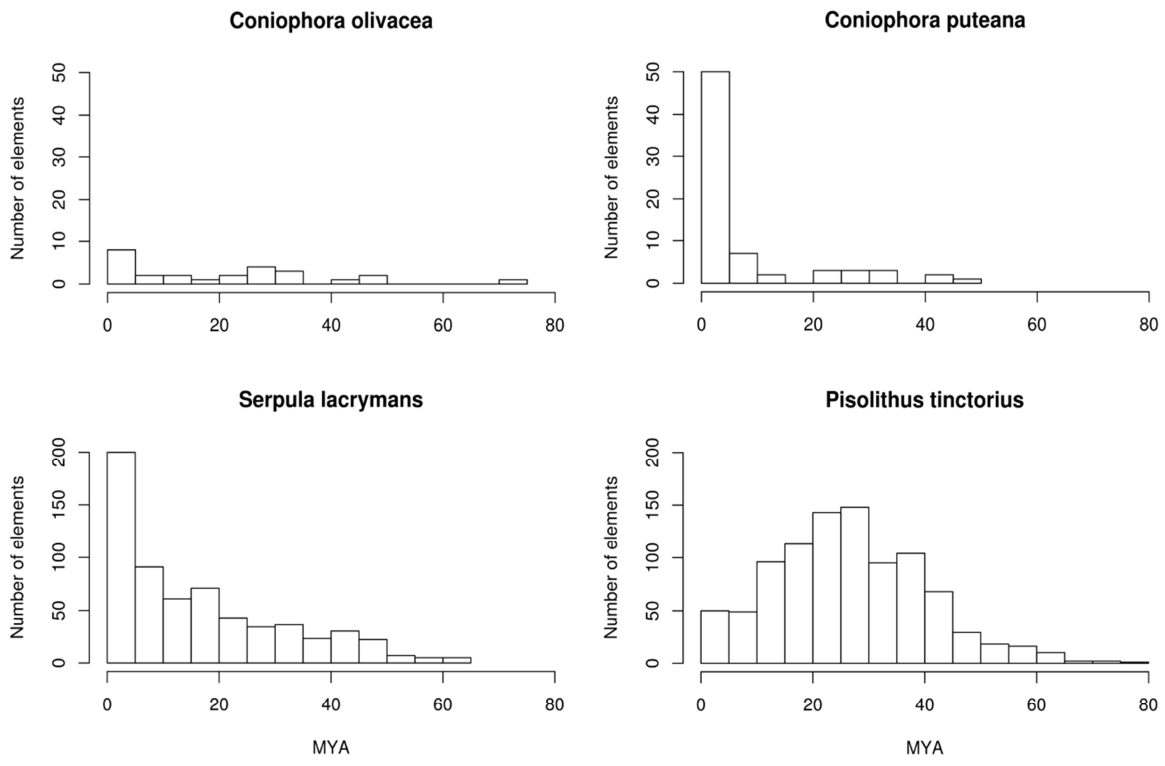

Figure 6. Estimated insertion age of the LTR-retrotransposons found in *C. olivacea, C. puteana, S. lacrymans* and *P. tinctorius.* MYA = million years ago.

4.4. Discussion

Genomic and proteomic characteristics of *C. olivacea*

We report 39.07 Mb assembly and annotation of brown rot basidiomycete *C. olivacea*. In terms of genome size, this species is slightly smaller than its sister species *C. puteana*, but it falls in the range of other brown-rot basidiomycetes such as *Hydnomerulius pinastri* (38.3 Mb) and *Serpuyla lacrymans* (42.7 Mb). *C. olivacea* and *C. puteana* show macrosynteny and good conservation of protein-coding genes, although the former has up to 1,352 additional orphan genes—most of these are supported by structure and RNA evidence (i.e., no homology to any other known gene).

In this sense, the higher number of annotated genes in *C. olivacea* relative to *C. puteana* is probably related to the higher amount of assembled RNA contigs used to assist the annotation of the former (resulting from the higher RNAseq depth). The presence of about 10% of orphan genes is common in fungal genomes, and these genes often lack an *in silico* functional annotation as we found for *C. olivacea* (Grandaubert et al. 2015; Nagy et al. 2015).

Wood-decaying species require a complex enzymatic machinery to degrade lignin and to obtain nutrients. According to the CAZy enzymes identified in the genome, the *C. olivacea* proteome carries the main signatures of canonical brown-rot: (i) it completely lacks Class II peroxidases—enzymes primarily involved in lignin degradation (Fernández-Fueyo et al. 2014), and (ii) it carries a reduced set of enzymes involved in degradation of crystalline cellulose. In fact, its profile is very similar to that of *C. puteana*. It displays only minor differences in several enzyme groups. As previously seen in other wood-degrading fungi, the *in silico* secretome of *C. olivacea* is enriched in functions related to lignocellulose degradation (Alfaro et al. 2016). Our analysis showed that most intracellular and secreted proteins were members of multi-gene families of diverse size originating from gene duplications. The number of gene families that could not be functionally annotated by standard similarity-based methods was remarkable. This is common in most fungal genomes.

To overcome this drawback, we used an alternative approach that combines similarity with structural information (Phyre-2). We then assigned a putative function to two multi-gene families conserved across the basidiomycete phylogeny but for which a putative function had not been previously proposed. Of especial interest is the newly identified family of putative copper-dependent lytic polysaccharide monooxygenases (AA9, LPMO). The LPMOs are recently discovered enzymes used by microbes to digest crystalline polysaccharides (Vaaje-Kolstad et al. 2010). They increase the saccharification yield of commercial enzyme cocktails (Müller et al. 2015). Despite the promising results obtained *in silico*, experimental assays will be necessary to confirm the function of the members of this newly described family.

Impact of TEs in the evolution of Boletales genomes

The results of TE annotation in the four Boletales showed how different patterns of LTR-retrotransposon amplifications have shaped the architecture of their genomes. The expansion of LTR/Gypsy retrotransposons belonging to the *Chromoviridae* clade occurred exclusively in the species with large genomes, whereas the smaller genomes have a small amount of these families (three families in *C. olivacea* and *C. puteana).* Chromoviruses are the most common LTR-retrotransposons in fungi (Muszewska et al. 2011), and the key to their success might be the presence of a chromo-integrase, which is thought to guide the integration of these elements into heterochromatic regions (Gao et al. 2008).

Heterochromatin is gene-poor, and it is silenced by epigenetic mechanisms such as DNA methylation and RNAi (Lippman and Martienssen 2004). Thus, integration of these elements in such regions would allow them to skip purifying selection and increase their probability to persist in the genome. In fact, this could be the reason for the longer prevalence of *Gypsy* over *Copia* LTR-retrotransposons in most fungal species—the latter tends to integrate at random locations including euchromatic regions where transposon fixation is more difficult (Pereira 2004).

The LTR-retrotransposon amplification bursts of the Boletales indicate that elements from both *Coniophora* species are young and thus putatively active. The profile of *S. lacrymans* also indicates a very strong activity of young copies with a progressive decrease in the amplification signals of older elements. The profile of *P. tinctorius* is intriguing—this ectomycorrhizal (ECM) species undergoes a massive expansion of LTR-retrotransposons in the Gypsy superfamily (similar to that found for other symbiotic species in Agaricomycotina (Labbe et al. 2012; Hess et al. 2014)); however, the majority of elements are very old (20-40 MYA) and still carry structural and coding domains necessary for transposition. This might be linked to a weaker genome defense activity or to a potential benefit of carrying an important transposon load although this remains to be demonstrated.

The phylogeny of Gypsy reverse-transcriptases suggests that many *P. tinctorius* families are distantly related to the other three species. This means that the extent of the amplification burst might be partially explained by the presence of a very active ancient family that exclusively colonized the genome of *P. tinctorius*. In addition, this finding shows that *S. lacrymans*, *C. olivacea* and *C. puteana* are currently in a period of genome expansion, but such a process is slowing in *P. tinctorius*.

Interestingly, in the latter case, the LTR-mediated genome amplification coincides with the estimated origins of ECM symbiosis in boletales (Kohler et al. 2015). Of the four Class I TE orders found here, only the LTR elements were present in the four species. We hypothesize that the most plausible scenario is that the elements from the other three orders (DIRS, LINE, and PLE) were lost by random drift in some of the species. Previous studies have shown that these orders tend to be present in low amount in other basidiomycetes (Castanera et al. 2016). Alternatively, they might be present in some genomes but in the form of very ancient and degenerated copies that are not detectable. Similarly, this patchy distribution was also found in class II elements—helitrons were absent in the *Coniophora* genus and present in the remaining two species.

4.5. References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–10. doi: 10.1016/S0022-2836(05)80360-2

Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. Nucleic Acids Res 19 Suppl:2247–9.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) The Pfam Protein Families Database. Nucleic Acids Res 30:276–280. doi: 10.1093/nar/30.1.276

Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome Res 10:547–8. doi: 10.1101/gr.10.4.547

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes databse (CAZY): an expert resource for Glycogenomics. doi: 10.1093/nar/gkn663

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. doi: 10.1093/bioinformatics/btp348

Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev I V., Stajich JE, Ramírez L (2016) Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet. doi: 10.1371/journal.pgen.1006108

Castanera R, Omarini A, Santoyo F, Pérez G, Pisabarro AGAG, Ramírez L (2013) Non-additive transcriptional profiles underlie dikaryotic superiority in *Pleurotus ostreatus* laccase activity. PLoS One 8:e73282. doi: 10.1371/journal.pone.0073282

Cortázar AR, Aransay AM, Alfaro M, Oguiza JA, Lavín JL (2014) SECRETOOL: integrated secretome analysis tool for fungi. Amino Acids 46:471–3. doi: 10.1007/s00726-013-1649-z

Dhillon B, Gill N, Hamelin RC, Goodwin SB (2014) The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. BMC Genomics 15:1132. doi: 10.1186/1471-2164-15-1132

Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, Blumentritt M, Coutinho PM, Cullen D, De Vries RP, Gathman A, Goodell B, Henrissat B, Ihrmark K, Kauserud H, Kohler A, LaButti K, Lapidus A, Lavin JL, Lee Y-H, Lindquist E, Lilly W, Lucas S, Morin E, Murat C, Oguiza JA, Park J, Pisabarro AG, Riley R, Rosling A, Salamov A, Schmidt O, Schmutz J, Skrede I, Stenlid J, Wiebenga A, Xie X, Kües U, Hibbett DS, Hoffmeister D, Högberg N, Martin F, Grigoriev I V, Watkinson SC (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science (80-) 333:762–765. doi: 10.1126/science.1205411

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi: 10.1093/nar/gkh340

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18. doi: 10.1186/1471-2105-9-18

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–84. doi: 10.1093/ nar/30.7.1575

Fernández-Fueyo E, Ruiz-Dueñas FJ, Martínez MJ, Romero A, Hammel KE, Medrano FJ, Martínez AT (2014) Ligninolytic peroxidase genes in the oyster mushroom genome: heterologous expression, molecular structure, catalytic and stability properties, and lignin-degrading ability. Biotechnol Biofuels 7:2. doi: 10.1186/1754-6834-7-2

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, De Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, John FS, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev I V, Hibbett DS (2012) The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science (80-) 336:1715–1719. doi: 10.1126/science.1221748

Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526. doi: 10.1371/journal.pone.0016526

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 18:359–369. doi: 10.1101/gr.7146408

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci 108:1513–1518. doi: 10.1073/pnas.1017351108

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–52. doi: 10.1038/nbt.1883

Hess J, Skrede I, Wolfe BE, LaButti K, Ohm RA, Grigoriev I V, Pringle A (2014) Transposable element dynamics among asymbiotic and ectomycorrhizal *Amanita* fungi. Genome Biol Evol 6:1564–1578. doi: 10.1093/gbe/evu121

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lücking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kõljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schüßler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N (2007) A higher-level phylogenetic classification of the Fungi. Mycol Res 111:509–547. doi: 10.1016/j.mycres.2007.03.004

Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11:431. doi: 10.1186/1471-2105-11-431

Jurka J (2000) Repbase Update - a database and an electronic journal of repetitive elements. Trends Genet 16:418–420. doi: 10.1016/S0168-9525(00)02093-X

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–66. doi: 10.1093/nar/gkf436

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858. doi: 10.1038/nprot.2015.053

Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–64. doi: 10.1101/gr.229202. Article published online before March 2002

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canback B, Choi C, Cichocki N, Clum A, Colpaert J, Copeland A, Costa MD, Dore J, Floudas D, Gay G, Girlanda M, Henrissat B, Herrmann S, Hess J, Hogberg N, Johansson T, Khouja HR, LaButti K, Lahrmann U, Levasseur A, Lindquist EA, Lipzen A, Marmeisse R, Martino E, Murat C, Ngan CY, Nehls U, Plett JM, Pringle A, Ohm RA, Perotto S, Peter M, Riley R, Rineau F, Ruytinx J, Salamov A, Shah F, Sun H, Tarkka M, Tritt A, Veneault-Fourrey C, Zuccaro A, Tunlid A, Grigoriev I V, Hibbett DS, Martin F (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet 47:410–415. doi: 10.1038/ng.3223

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. doi: 10.1186/gb-2004-5-2-r12

Larraya LM, Perez G, Penas MM, Baars JJP, Mikosch TSP, Pisabarro AG, Ramirez L (1999) Molecular Karyotype of the White Rot Fungus *Pleurotus ostreatus*. Appl Envir Microbiol 65:3413–3417.

Lippman Z, Martienssen R (2004) The role of RNA interference in heterochromatic silencing. Nature 431:364–70. doi: 10.1038/nature02875

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–64.

Martin F, Kohler A, Murat C, Veneault-Fourrey C, Hibbett DS (2016) Unearthing the roots of ectomycorrhizal symbioses. Nat Rev Microbiol 14:760–773. doi: 10.1038/nrmicro.2016.149

Melén K, Krogh A, von Heijne G (2003) Reliability measures for membrane protein topology prediction algorithms. J Mol Biol 327:735–44.

Müller G, Várnai A, Johansen KS, Eijsink VGH, Horn SJ (2015) Harnessing the potential of LPMO-containing cellulase cocktails poses new demands on processing conditions. Biotechnol Biofuels 8:187. doi: 10.1186/s13068-015-0376-y

Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR retrotransposons in fungi. PLoS One 6:e29425. doi: 10.1371/journal.pone.0029425

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–5. doi: 10.1093/bioinformatics/btt509

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6. doi: 10.1093/protein/10.1.1

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–3. doi: 10.1093/bioinformatics/btt054

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27:29–34. doi: 10.1093/nar/27.1.29

Pereira V (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 5:R79. doi: 10.1186/gb-2004-5-10-r79

Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. Bioinformatics 21:I351–I358. doi: 10.1093/bioinformatics/bti1018

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. doi: 10.1093/molbev/msp077

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1:166–175. doi: 10.1371/journal.pcbi.0010022

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–20. doi: 10.1093/nar/gki442

Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, Levasseur A, Lombard V, Morin E, Otillar R, Lindquist EA, Sun H, LaButti KM, Schmutz J, Jabbour D, Luo H, Baker SE, Pisabarro AG, Walton JD, Blanchette RA, Henrissat B, Martin F, Cullen D, Hibbett DS, Grigoriev I V (2014) Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A 111:9923–9928. doi: 10.1073/pnas.1400592111

Salamov AA, Solovyev V V (2000) Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10:516–22. doi: 10.1101/gr.10.4.516

SanMiguel P, Gaut BS, Tikhonov a, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45. doi: 10.1038/1695

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. doi: 10.1186/1471-2105-6-31

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–3. doi: 10.1093/bioinformatics/btu033

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–77. doi: 10.1080/10635150701472164

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov A V, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. doi: 10.1186/1471-2105-4-41

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res 18:1979–90. doi: 10.1101/gr.081612.108

Vaaje-Kolstad G, Westereng B, Horn SJ, Liu Z, Zhai H, Sorlie M, Eijsink VGH (2010) An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides. Science (80-) 330:219–222. doi: 10.1126/science.1192231

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35:D5–12. doi: 10.1093/nar/gkl1031

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi: 10.1038/nrg2165

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–9. doi: 10.1101/gr.074492.107

Chapter V: Biology, dynamics and applications of transposable elements in basidiomycete fungi (review & general discussion)

Sections of this chapter have been published as: Castanera R, Borgognone A, Pisabarro AG, Ramírez L (2017) Biology, dynamics, and applications of transposable elements in basidiomycete fungi. Appl Microbiol Biotechnol 1–14. doi: 10.1007/s00253-017-8097-8. Impact Factor: 3.37

5.1. Diving into TEs: Considerations abount annotation

Basidiomycete fungi encompass species with highly diverse lifestyles, such as plant pathogens, saprophytes, mycorrhizas, animal pathogens and endophytes, among others. An important number of species from these groups are being sequenced as part of multiple large-scale genome projects. As a result, the amount of genomic information increases in a monthly (or even weekly) basis. These data have revealed very important hints in terms of coding features, such as expansions or contractions of specific gene families correlated to the proliferation of the different lifestyles (Floudas et al, 2012; Kohler et al, 2015). Another striking discovery is the high variability in genome sizes (Nemri et al. 2014; Dutheil et al. 2016). In other eukaryotes, sharp increments in genome sizes are usually linked to the expansion of non-coding DNA such mobile elements (Lynch 2007). TE content is extremely variable among eukaryotes (Canapa et al, 2016; Chénais et al. 2012) and fungi are not an exception. Within the fungal kingdom there are species practically free of TEs such as *Pseudozyma hubiensis* or *Mixia osmundae* (described in Chapter III), and others with at least 90% of their genome composed by repeats, such as *Blumeria graminis* (Wicker et al. 2013). The main difficulty when attempting to compare TE abundance and distribution in basidiomycetes comes from variability introduced by TE annotation procedures. Due to the repetitive nature of these mobile elements, their annotation in assembled genomes is a complex task that requires a combination of multiple dedicated tools (Lerat 2010). In addition, the assembly quality critically influences TE detection. Draft genomes often underestimate their number in comparison to what it is found in complete, telomere-to-telomere assemblies. Clear examples of these are the difficulties found to annotate complete helitrons and LTR-retrotransposons in PC9 (genome draft) vs PC15 assemblies (complete assembly), as shown in Chapters II and III. Two main approaches can be used to identify TEs: *de novo* and homology-based approaches. High quality TE annotations require the use of several *de novo* methods, such as those based on element repetitiveness (genome self-comparison) or on structural features of specific TE superfamilies (i.e., LTR: Long Terminal Repeats or TIR: Terminal Inverted Repeats), followed by manual curation. In fact, the majority of basidiomycete TE annotations published until now have been performed using a combination of *de novo* and homology-based approaches using different software programs, parameters, and cutoffs (Table 1).

Table 1. Summary of TE content and genomic features of 65 basidiomycetes (adapted from Castanera et al, 2017). Information about the phylogeny, lifestyle, genomic characteristics, and a brief description of the approach used for the annotation of TEs is shown for each species.

Species	Lifestyle	Genome size (Mb)	Genes	TE content (%)			Total TE	Methodology [&]	Source
				Class I	Class II	Unknown			
Ustilaginomycotina									
Pseudozyma antarctica	PA / SAP yeast	18.1	6,640		0.01	0.09	0.1	RepeatScout + RECON + LTRharvest	Morita T et al., 2013 Present study (Chapter III)
Pseudozyma hubeiensis	PA / SAP yeast	18.4	7,472		0.01	0.11	0.12	RepeatScout + RECON + LTRharvest	Konishi M et al., 2013 Present study (Chapter III)
Sporisorium reilianum	PP / SAP yeast	18.4	6,648	0.18	0.05	0.01	0.24	RepeatScout + Repbase	Schirawski J et al., 2010 Laurie et al, 2012
Malassezia globosa	AP / SAP yeast	8.9	4,286	0.39	0.16		0.55	RepeatScout + Repbase	Xu J et al., 2007 Dutheil et al, 2016
Tilletiaria anomala	AP / SAP yeast	18.7	6,810				1	RepeatScout + Repbase	Toome et al, 2014
Ustilago maydis	PP / SAP yeast	19.8	6,902	1.61	0.1	0.07	1.78	RepeatScout + Repbase	Kamper J et al., 2006 Laurie et al, 2012
Ustilago hordei	PP / SAP yeast	21.2	7,113	7.23	0.44	0.23	7.9	RepeatScout + Repbase	Laurie et al, 2012
Pucciniomycotina									
Mixia osmundae	PP / SAP yeast	13.63	6,903				**0.15** / 1.5	**RepeatScout + Repbase**	**Toome et al, 2014** / Present study (Chapter III)
Rhodotorula graminis	E yeast	21.03	7,283				3.63	RepeatScout + Repbase	Firrincieli et al, 2015
Microbotryum lychnidis-dioicae	PP (OB)	26.1	7,364	5.01	3.2	5.85	14.06	TEdenovo + TEannot (REPET pipeline)	Perlin et al, 2015
Puccinia striiformis	PP (OB)	64.8	18,021	8.2	8.2	1.3	17.7	Repbase + Libraries from Duplessis et al, (2011)	Cantu et al, 2011
Puccinia graminis	PP (OB)	88.6	17,773	13.44	11.73	18.6	**43.77** / 43.3	**TEdenovo + TEannot (REPET pipeline)**	**Duplessis et al, 2011** / Present study (Chapter III)
Melampsora larici-populina	PP (OB)	101.1	16,339	11.6	15.2	18.2	45	TEdenovo + TEannot (REPET pipeline)	Duplessis et al, 2011
Melampsora lini	PP (OB)	189.5	16,271	24.16	7.53	13.51	45.2	EvidenceModeler (Blast2go)	Nemri A et al, 2014
Agaricomycotina									
Schizopora paradoxa	WR	44.4	17,098				0.6	RepeatScout + Repbase	Min et al, 2015
Hebeloma cylindrosporum	EM	38.2	15,382				0.8	RepeatScout + Repbase	Kohler et al, 2015
Wallemia sebi	XF	9.8	5,284				0.8	Repbase +RepeatMasker	Padamsee et al, 2012
Hypholoma sublateritium	WR	48	17,911				1.1	RepeatScout + Repbase	Kohler et al, 2015
Ceriporiopsis subvermispora	WR	38.97	12,125				1.2	Repbase +RepeatMasker	Fernandez-Fueyo et al, 2012

Bjerkandera adusta	WR/ PP	42.7	15,473				1.34	RepeatScout + Repbase	Binder et al, 2013
Plicaturopsis crispa	WR	34.5	13,626				1.5	RepeatScout + Repbase	Kohler et al, 2015
Wallemia ichthyophaga	HF	9.6	4,884				1.67	Repbase +RepeatMasker	Zajc et al, 2013
Coniophora olivacea	BR	39.07	14,928	1.08	1	0.07	2.15	TEdenovo + TEannot (REPET pipeline)	Present study (Chapter IV)
Hydnomerulius pinastri	BR	38.3	13,270				2.3	RepeatScout + Repbase	Kohler et al, 2015
Suillus luteus	EM	37	18,316				2.4	RepeatScout + Repbase	Kohler et al, 2015
Trichosporon oleaginosus	O yeast	19.8	8,322				2.85	RepeatScout + Repbase	Kourist et al, 2015
Stereum hirsutum	WR / PP	45.64	14,072	0.41	0	2.68	3.09	RepeatScout + LTR_STRUC	Floudas et al, 2012
Sebacina vermifera	OS	38.1	15,312				3.9	RepeatScout + Repbase	Kohler et al, 2015
Trametes versicolor	WR	42.88	14,296	0.68	0.14	3.35	4.17	RepeatScout + LTR_STRUC	Floudas et al, 2012
Coniophora puteana	BR	41.86	13,761	1.18	0.08	2.93	**4.19** / 3.94	**RepeatScout + LTR_STRUC**	**Floudas et al, 2012 /** Present study (Chapter IV)
Phlebia brevispora	WR	49.96	16,170				4.53	RepeatScout + Repbase	Binder et al, 2013
Volvariella volvacea (PS)	SAP	52.4		3.54	0.55	0.53	4.62	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Piriformospora indica	E	24.98	11,769	0.56		4.12	4.68	RepeatScout + LTR_STRUC	Zuccaro A et al, 2011
Amanita inopinata	SAP	22.1		4.29	0.07	0.44	4.8	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Dacryopinax primogenitus	BR	27.6	10,242	1.56	0.64	2.61	4.81	RepeatScout + LTR_STRUC	Floudas et al, 2012
Tulasnella calospora	OS	62.4	19,659				4.9	RepeatScout + Repbase	Kohler et al, 2015
Pleurotus ostreatus (PC9)	WR	35.6	12,206	2.39	0.11	2.41	4.91	RepeatScout + RECON + LTRharvest	Present study (Chapter III)
Punctularia strigoso-zonata	WR	33.07	11,538	2.22	0	2.97	5.19	RepeatScout + LTR_STRUC	Floudas et al, 2012
Piloderma croceum	EM	59.3	21,583				5.9	RepeatScout + Repbase	Kohler et al, 2015
Cryptococcus neoformans var. grubii	AP	18.87	6,967	3.43	0.54	1.94	5.91	RepeatScout + RECON + LTRharvest	Janbon et al. 2014 Present study (Chapter III)
Coprinopsis cinerea	SAP	36.29	13,342	4.37	0.02	1.68	6.07	RepeatScout + LTR_STRUC	Stajich JE et al. 2010 Floudas et al, 2012
Suillus brevipes	EM	51.7	22,453				6.11	RepeatScout + Repbase	Branco S et al, 2015
Volvariella volvacea (V23)	SAP	35.7	11,084				6.18	RepeatScout	Bao D et al, 2013
Laccaria amethystina	EM	52.2	21,066				6.5	RepeatScout + Repbase	Kohler et al, 2015
Cryptococcus neoformans var neoformans	AP	19.05	6,475	4.48	0.78	1.38	6.64	RepeatScout + RECON + LTRharvest	Loftus et al 2005 Present study (Chapter III)
Ganoderma sp.	WR	43.3	16,113	5.42	1.67	0.6	7.69	TEdenovo + TEannot (REPET pipeline)	Chen et al, 2012
Gymnopus luxurians	SAP	66.3	22,057				7.7	RepeatScout + Repbase	Kohler et al, 2015
Amanita muscaria	EM	67.6		5.54	0.78	1.41	7.73	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Gloeophyllum trabeum	BR / PP	34.43	11,846	1.75	0.03	6.3	8.08	RepeatScout + LTR_STRUC	Floudas et al, 2012

Rhizoctonia solani	PP	39.8	13,964	4.13	0.09	4.17	8.39	RepeatScout +Repbase	Hane, 2014
Paxillus involutus	EM	58.3	17,968				8.4	RepeatScout + Repbase	Kohler et al, 2015
Amanita muscaria Koide	EM	40.7	18,153	6.49	0.96	1.46	8.91	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014 Kohler et al, 2015
Dichomitus squalens	WR	39.45	12,290	3.92	0.11	5.13	9.16	RepeatScout + LTR_STRUC	Floudas et al, 2012
Fomitopsis pinicola	BR	42.06	14,724	2.45	0.1	6.74	9.29	RepeatScout + LTR_STRUC	Floudas et al, 2012
Pleurotus ostreatus (PC15)	WR	34.3	12,330	5.88	0.28	3.79	9.95	RepeatScout + RECON + LTRharvest	Present study (Chapter III)
Phanerochaete chrysosporium	WR	35.14	11,777	6.55	0.25	3.38	10.18	RepeatScout + RECON + LTRharvest	Martinez et al. 2004 Present study (Chapter III)
Auricularia delicata	WR	69.05	23,577	2.63	0.27	7.55	10.45	RepeatScout + LTR_STRUC	Floudas et al, 2012
Schizophyllum commune	WR	38.5	13,210	4.66	0.62	5.51	10.79	RepeatScout + RECON	Ohm et al, 2010
Sphaerobolus stellatus	WR	176.4	35,274				10.9	RepeatScout + Repbase	Kohler et al, 2015
Amanita polypyramis	EM	23.5	-	11.22	0.18	0.23	11.63	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Agaricus bisporus var bisporus	SAP	30.23	10,438	7.86	0.99	3.58	12.43	RepeatScout + LTR_STRUC +tBLASTx	Foulongne-Oriol et al, 2013
Agaricus bisporus var. burnettii	SAP	32.6	11,289	7.54	0.25	6.82	14.61	RepeatScout + LTR_STRUC +tBLASTx	Foulongne-Oriol et al, 2013
Pisolithus microcarpus	EM	53	21,064				14.8	RepeatScout + Repbase	Kohler et al, 2015
Heterobasidion annosum	WR / PP	33.64	11,464	9.78	0.32	5.86	15.96	RepeatScout + LTR_STRUC	Olson A et al., 2012 Floudas et al, 2012
Phanerochaete carnosa	WR / PP	46.3	13,937	6.41	0.58	11.42	18.41	RepeatScout + RECON + LTRharvest	Suzuki et al. 2012 Present study
Tremella mesenterica	WR / MP	27.98	8,313	14.48	1.24	3.45	19.17	RepeatScout + LTR_STRUC	Floudas et al, 2012
Scleroderma citrinum	EM	56.1	21,012				22.3	RepeatScout + Repbase	Kohler et al, 2015
Laccaria bicolor	EM	60.7	20,614	10.33	4.98	8.88	**24.19** / 37.9	**TEdenovo + TEannot (REPET pipeline)**	Martin et al. 2008 **Labbé et al, 2012** / Present study (Chapter III)
Paxillus rubicundulus	EM	53	22,065				25.7	RepeatScout + Repbase	Kohler et al, 2015
Amanita thiersii	SAP	33.7	10,354	25.8	0.05	0.31	26.16	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Wolfiporia cocos	BR	48.24	12,746	5.31	0.93	21.47	27.71	RepeatScout + LTR_STRUC	Floudas et al, 2012
Pisolithus tinctorius	EM	71	22,701				**29.8** / 41.17	**RepeatScout + Repbase**	**Kohler et al, 2015**/ Present study (Chapter IV)
Serpula lacrymans (S7.3)	BR	47.04	14,495	24.33	1.68	7.04	**33.05** /29.45	**RepeatScout + RECON + LTRharvest**	**Present study Chapter III** / Chapter IV
Amanita brunnescens	EM	57.6	-	32.77	0.99	1.94	35.7	TEdenovo + TEannot (REPET pipeline)	Hess et al, 2014
Serpula lacrymans (S7.9)	BR	42.73	12,789	27.78	2.05	8.29	38.12	RepeatScout + RECON + LTRharvest	Present study (Chapter III)
Fomitiporia mediterranea	WR / PP	56.77	11,333	27.58	5.48	8.36	41.42	RepeatScout + LTR_STRUC	Floudas et al, 2012

SAP - Saprophytic; PP - Plant pathogen; (OB) - obligate biotroph; PA - Plant-associated (non-pathogenic); AP - Animal pathogen; E – Endophytic; SAP yeast -Saprophytic yeast; WR - White rot; BR - Brown rot; MP - Mycoparasite; EM - Ectomycorrhizal fungi; OS- Orchid symbiont; O - Oleaginous ; XF - Xerophilic Fungi; HF - Halophilic fungi.

& Description of software/database used for detection of TEs.

Bold and regular font is used in cells with more than one TE content, methodology or source, to differentiate the results of the different approaches

RepeatScout software (Price et al. 2005) seems to be the preferred method for *de novo* identification of repeats and is often followed by structure-based methods, such as LTRharvest (Ellinghaus et al. 2008) or LTR_STRUCT (McCarthy and McDonald 2003) to detect full-length LTR-retrotransposons. Nevertheless, the accurate annotation of certain TE families requires manual curation, as their particular features are very difficult to detect by automatic tools. We have shown in Chapter II how Helsearch (a program specifically designed to identify helitrons in plant genomes) yielded multiple false positives, and skipped the detection of a very important family (HELPO2) due a small variant in the 3` terminal structure. This is also applicable to other complex transposons such as TRIM and LARD retrotransposons, which can be easily misannotated by automatic tools due to the relatively vague features that characterize them. Classification of TEs identified *de novo* is usually performed using similarity searches against databases such Repbase (Jurka et al. 2005) or PFAM (Finn et al. 2014) to identify TE-specific domains. These can be complemented with more sophisticated software that is able to detect structural features such TIRs or LTRs (Hoede et al. 2014). Often, custom libraries are constructed including fungal reference sequences available in Repbase. The construction of such species-specific library is essential for an accurate TE annotation, as an important fraction of TEs are undetectable by homology-based programs when they are fed with libraries from other organisms. Also, combining several strategies leads to the most reliable results. In fact, the results obtained for some species analyzed by the pipeline described in Chapter III and by REPET pipeline (Flutre et al. 2011) are quite similar (ie, *Puccinia graminis*, *Serpula lacrymans*), despite the detection programs used in the two pipelines are different. Currently, the most commonly used tool to perform the final TE annotation during genome assembly is RepeatMasker. This process must be followed by a de-fragmentation step to join TE fragments into full-length copies (Flutre et al. 2011; Castanera et al. 2016).

5.2. A snapshot of the distribution of TEs in Basidiomycetes

The phylum *Basidiomycota* includes three clades: *Pucciniomycotina*, *Ustilaginomycotina*, and *Agaricomycotina* (Hibbett et al. 2007). This latter clade has the most sequenced and publicly available genomes. Despite the TE content in basidiomycetes ranging from 0.1% to 45.2% of their genomes, TE abundance in most species is low (average of 11%, Fig. 1A, Table 1). Both class I and class II TEs populate basidiomycete genomes, but the relative abundance of each TE class varies within the three subphyla. For *Agaricomycotina*, class I elements are clearly dominant, whereas class II elements are constrained and usually do not comprise more than

1% of the genome. By contrast, species from the *Pucciniomycotina* clade harbor a much more balanced ratio of class I/class II elements. Interestingly, both classes show important expansions in rust plant pathogens (Duplessis et al. 2011). Within the *Ustilaginomycotina*, there is not a clear trend for TE abundance besides an important expansion of class I elements in *Ustilago hordei* (Laurie et al. 2012). Class I elements comprise 7.23% of the *Ustilago hordei* genome*;* in contrast, TEs are practically absent in other sequenced species of this clade. Among the vast diversity of class I TEs, the LTR-retrotransposon superfamilies *Gypsy* and *Copia* usually dominate the landscape of basidiomycete genomes (Floudas et al. 2012; Labbe et al. 2012; Foulongne-Oriol et al. 2013; Castanera et al. 2016). Other class I elements such as tyrosine recombinase (YR) retrotransposons and non-LTR retrotransposons like *Tad1* and *L1* are commonly found populating genomes of *Pucciniomycotina* and *Agaricomycotina*, but in lower amounts (Muszewska et al. 2013; Castanera et al. 2016).

Regarding class II transposons, Helitrons and several cut-and-paste TIR superfamilies such *Tc1-Mariner*, *Harbinger*, or *En/Spm* have been identified in several basidiomycetes (Hood 2005; Floudas et al. 2012; Castanera et al. 2014; Castanera et al. 2016), although these superfamilies are not highly abundant. As in mammals and plants, LTR-retrotransposons tend to accumulate in clusters in the basidiomycete genomes, including *Laccaria bicolor* (Labbe et al. 2012), *Pleurotus ostreatus* (Castanera et al. 2016), *Agaricus bisporus* (Sonnenberg et al. 2016) and *Coprinopsis cinerea*, where these clusters match with the cytological centromeres (Stajich et al. 2010). In addition, estimations of LTR amplification bursts indicate recent expansions of these elements in Basidiomycetes (Labbe et al. 2012; Foulongne-Oriol et al. 2013; Castanera et al. 2016). The replicative mechanism of class I elements ensures an efficient increase in their copy number, which may be the key for their proliferative success in comparison to class II transposons. In this sense, the unique expansion of class II elements described in rust fungi (Duplessis et al. 2011; Nemri et al. 2014) challenges that rule and our understanding of TE dynamics. Also, it is possible that the abundance of class II transposons is underestimated by basidiomycete TE annotations because of their low copy number which makes them difficult to identify by *de novo* approaches. Another important aspect of basidiomycete TEs is the low ratio of autonomous *vs.* non-autonomous degenerated elements, the latter usually outnumber the former by a factor of 10 to 100 (Labbe et al. 2012, Chapter III) although this ratio can vary depending the quality of the assembly and the approach used for annotation.

5.3. Influence of TEs on Basidiomycetes genome size

It is generally accepted that eukaryotic genome size is deeply affected by the dynamics of

repetitive DNA elements such as TEs, rDNAs, tandem duplications, DNA satellite, etc. This relationship has been widely explored in plants and animals, where an important part of the variability in genome size is explained by expansions and contractions of different transposon groups (Lee and Kim 2014; Canapa et al. 2016). Previous studies in fungi have shown massive amplifications of TEs in the largest genomes (Martin et al. 2010; Spanu et al. 2010), whereas the smaller ones are almost depleted of repetitive DNA (Toome et al. 2014; Dutheil et al. 2016). In the species surveyed in this chapter, an important amount (41%) have a low TE content (0-5%), whereas very few genomes have high TE content (Table 1, Fig. 1A). In addition, the correlation between genome size and gene content is higher than between genome size and TE content (Spearman's correlation p-value < 0.01; $\rho = 0.90$ vs $\rho = 0.55$, Fig. 1D). Further, every basidiomycete subphylum shows a particular behavior in terms of genome size versus TE coverage or gene content (Fig. 1D and 1E, respectively). Species from *Ustilaginomycotina* have very compact genomes with low gene numbers and a TE content that ranges from 0.1 to 7.9% of the genome (Laurie et al. 2012; Dutheil et al. 2016; Castanera et al. 2016). In *Pucciniomycotina*, two main groups are observed. The first group comprises the rust species, with large genomes that display important TE expansions (Duplessis et al. 2011). The second group contains yeasts, such as *Rhodotorula graminis* and *Mixia osmundae*, which show a particular genomic contraction along with low gene and repetitive content (Toome et al. 2014; Firrincieli et al. 2015). The strong impact of non-coding DNA on genomic expansions in *Pucciniomycotina* species is easily observed when comparing TE content versus genome size (Fig. 1D) and gene number versus genome size (Fig. 1E). In *Pucciniomycotina*, the trend with gene number plateaus at approximately 17,000, deviating from the ascending norm of the rest of the *Agaricomycotina* species. Interestingly, the anther-smut fungus *Microbotryum lychnidis-dioicae* remains at an intermediate point. Its gene content is similar to that of the yeast group (7,364 genes), but its TE content is closer to that of the rusts (14.6%) (Perlin et al. 2015). In this respect, *Agaricomycotina* species show a paradoxical behavior; the genome size of most of them is distributed in a relatively narrow range (40 to 60 Mb) and the correlation between gene content and genome size is high. Nevertheles, the relationship between TE abundance and genome size is unclear. In fact, some species annotated with the same pipeline such as *Amanita brunnescens* and *Volvariella volvacea* have similar genome sizes but display large differences in TE content (Hess et al. 2014) (Table 1). The explanation for this paradox may reside in the differential presence of ancient relic TE, i.e., elements that usually lack conserved domains, structural features, or open reading frames, due to the accumulation of mutations. They have been referred to previously as genomic dark matter and are very difficult to detect and classify with the

canonical bioinformatics tools.

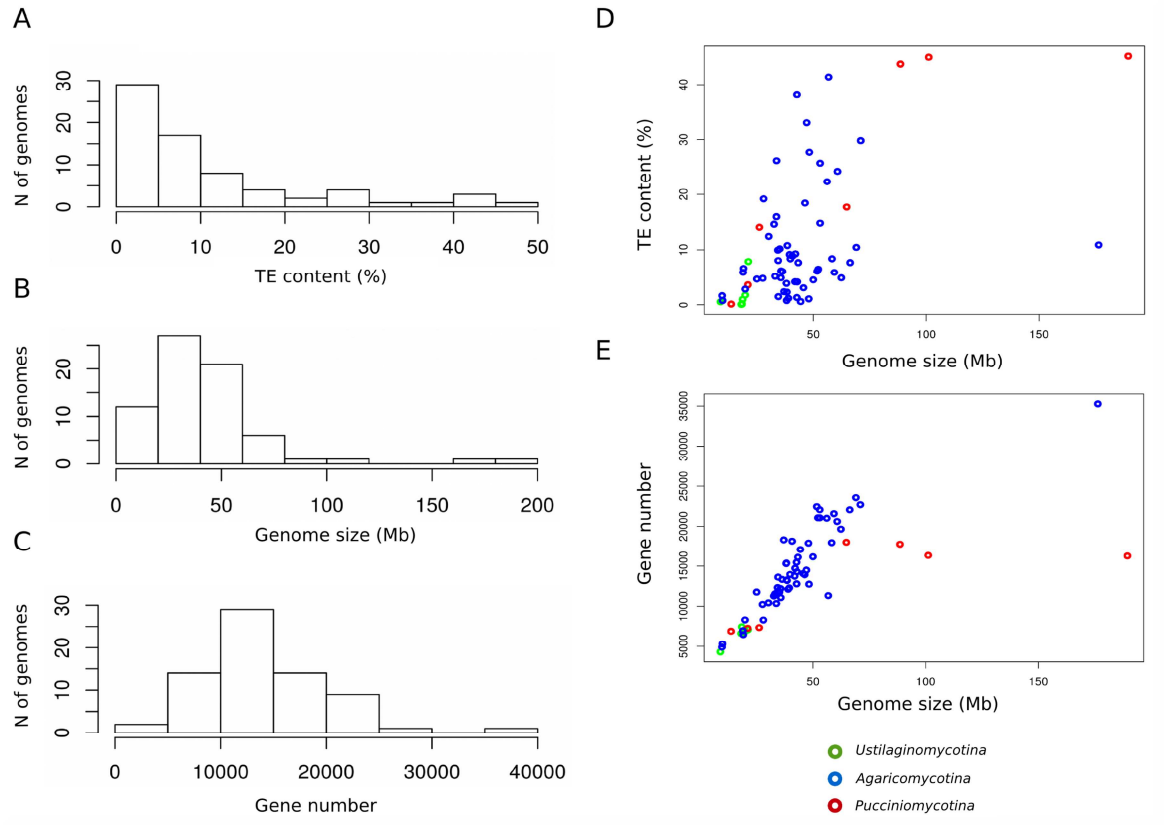

Figure 1. TE content and genomic features of Basidiomycetes. Distribution of TE content (A), genome size (B), and gene number (C) of species surveyed in this review. Scatterplots show the relationship between TE content and genome size (D) as well as gene content and genome size (E).

5.4. Impact of TEs on genomic architecture and functionality

According to previous research, there are numerous ways by which TEs can drive permanent modifications in fungal host genomes. TEs promote rearrangements, insertions, excisions, or create new alleles (Daboussi and Capy 2003). The result of most TE activity is likely neutral in basidiomycetes, as they often insert at intergenic regions. Nevertheless, TE insertions near or into genes can lead to gene inactivation as has been described for *lip12*, a gene that encodes a lignin peroxidase in *P. chrysosporium* (Gaskell et al. 1995). Interestingly, TE activity can

also be beneficial for certain species such as some filamentous plant pathogens (Raffaele and Kamoun 2012). For example, in *Ustilago hordei*, a TE insertion at a dominant avirulent locus offers the fungus the opportunity to overcome plant resistance by avoiding host recognition (Ali et al. 2014). Despite effects that TE insertions or TE-mediated rearrangements produce in the genome, TE repeats are probably essential for chromosomal architecture of most basidiomycetes. Results from studies in *C. cinerea* (Stajich et al. 2010) and *Cryptococcus neoformans* (Janbon et al. 2014) have shown enrichment of TEs at presumptive centromeres, suggesting that TEs play a role in centromere evolution in basidiomycetes similar to that found in plants and animals (Wong and Choo 2004). The limited amount of telomere-to-telomere genome assemblies hinders the study of TE distribution along the full chromosome, but it is common to find gene-poor, transposon-rich regions spanning 100 to 200 kb in some of the longest scaffolds of basidiomycete genome drafts; some of which probably coincide with centromeric regions. TE clusters are highly dynamic and can evolve rapidly as is shown by the recent amplification bursts of LTR-retrotransposons in *L. bicolor*, *A. bisporus*, *P. ostreatus,* and *Amanita* species (Labbe et al. 2012; Foulongne-Oriol et al. 2013; Hess et al. 2014; Castanera et al. 2016). Recent studies in *Agaricus bisporus* (Sonnenberg et al. 2016) and *P. ostreatus* showed that these TE clusters break the collinearity between close species and even between haplotypes. In addition, *P. ostreatus* genes present inside these TE clusters displayed lower expression than the average gene expression of the whole genome, suggesting that TEs also modulate gene expression of surrounding genes. In plants and animals, a handful of information has been described suggesting that TEs are powerful regulatory elements (Chuong et al. 2016). In fungi, the effect that TEs produce on genes´expression at a whole genome level is largely unknown, in part because the research community lacks of reference TE annotations. In Chapter III we hypothesize that the TE gene silencing effect found in *P. ostreatus* could be the result of the epigenetic inactivation of these TE knobs. In an ongoing study of our group, we have found that genes carrying TE insertions in 1kb upstreams/downstream windows (Fig 6, Chapter III) show higher levels of cytosine methylation than control genes, resulting from the extension of methylation from the adjacent TE (Borgognone et al, unpublished resuls).

5.5. TEs and basidiomycete lifestyles

One key question in the field of basidiomycete TEs is to understand their relationship with the host lifestyle. Certainly, there are important differences in the average transposon content of species displaying the main basidiomycete lifestyles (Fig 2). The most striking differences are

found between obligate biotrophic plant pathogens and yeasts (average TE content is 33.1% and 1.7%, respectively). The former group is represented by *Pucciniomycotina spp.*, suggesting that this phenomenon could be explained by their phylogenetic proximity. Nevertheless, there are clear incompatibilities with this hypothesis. For example, there are species in this subphylum that have a saprotrophic-yeast lifestyle and display very low TE content (i.e., *M. osmundae* and *R. graminis*, Table 1). In fact, the yeast group is comprised of species from three subphyla. Some of them are associated with plants (i.e., *Malassezia globosa*), and others can switch from the monokaryotic saprophytic yeast phase to a filamentous dikaryotic stage with parasitic lifestyle (i.e., smut fungi from the *Ustilago* genus). Nevertheless, they all have a small, compact, low-repetitive genome. In this sense, genomic data collected from unicellular and multicellular species suggests that there is a threshold genome size (around 10-20 Mb) below which mobile elements cannot be properly established in a population. Such a phenomenon could not be a consequence of immunity to TEs (i.e., inactivation through genome defense), but of the large effective population sizes of these species that difficult TE fixation (Lynch and Conery 2003). In basidiomycete yeasts with a parasitic lifestyle, secreted effectors (necessary for host interactions) are often arranged in rapidly evolving gene clusters associated with TEs, which play a key role in their evolution (Ali et al. 2014; Dutheil et al. 2016). Interestingly, it has been recently described that TE activity in smut pathogens may be restricted to such regions, where transposon-mediated variability can be beneficial (Dutheil et al. 2016). Large, repetitive genomes of obligate biotrophs such *Puccinia graminis* or *Melampsora larici-populina* also harbor hundreds of effector-like proteins called small-secreted proteins (SSP) that are likely involved in virulence. For example, some of the effectors described in *M. larici-populina* are scattered across the genome, but unlike it happened in the case of smut fungi, they were not found to be present in enriched TE regions (Duplessis et al. 2011). Although the reasons for these differences are not fully understood, the unique genomic features of filamentous obligate biotrophs have been suggested to be positive in a macroevolutionary context (Raffaele and Kamoun 2012), as large genomes are likely to adapt faster to novel conditions. For example, the plasticity of a TE-rich genome may confer an advantage to rust fungi when attempting to colonize new hosts. This is of great importance because these fungi cannot survive as free-living mycelia independently of a plant host. In contrast, less repetitive genomes of non-obligate biotroph smut fungi are not likely to undergo rapid evolution and this could contribute to their narrower host-range. Transposable elements also constitute an important portion of the genomes of basidiomycetes with saprophytic and symbiotic lifestyles. Floudas and colleagues (2012) highlighted that most *Agaricomycotina* species enriched for TEs are

involved in mutualistic or parasitic interactions with plants. The link between TE activity and plant-fungi associations was further explored in the *Amanita* genus (Hess et al. 2014), which contains both symbiotic (EM) (e.g., *A. brunnescens*, *A. polypyramis*, and *A. muscaria*) and asymbiotic (AS) (e.g., *A. inopinata* and *A. thiersii*) species. In this study, the authors found that despite TE content of EM species was high; such profile was not exclusive of the species in this group, as AS species showed variable TE content and activity. Nevertheless, two of the three EM species analyzed showed signs of recent amplification of different TE clades. Regarding ligninolytic fungi, the TE content of white rots (WR) is usually lower than that of brown rots (BR) (5.8% versus 17.1%, respectively). Nevertheless, in Chapter IV we have presented a BR species displaying very low TE content (*Coniophora olivacea*, 2.15%), as well as a BR with a very high TE load (*Serpula lacrymans*, 29.45%). Intriguingly, both species are phylogenetically close. The difference in TE content was mainly ascribed to LTR-retrotransposons, a TE group that is currently undergoing an expansion burst in both genomes. Also, is noteworthy to mention that some WR species with a pathogenic lifestyle such *Phanerochaete carnosa* or *Fomitiporia mediterranea* have experienced very important TE expansions (e.g., *Gypsy* LTR-retrotransposons), leading to a much higher TE abundance (18.4 % and 41.4 %, respectively) (Floudas et al. 2012; Castanera et al. 2016). The influence of TE content on WR and BR lifestyles has not been studied in depth. Nevertheless, a study described the effects of TE inactivation of a lignin peroxidase gene in *P. chrysosporium* (Gaskell et al. 1995). In Chapter III, we describe a TE-mediated silencing in *P. ostreatus* clusters carrying Carbohydrate Active Enzymes (CAZY). These enzymes play an important role on plant cell wall breakdown, and thus they are essential for ligninolytic fungi. These results suggest that TE activity does not convey a benefit to this group of fungi.

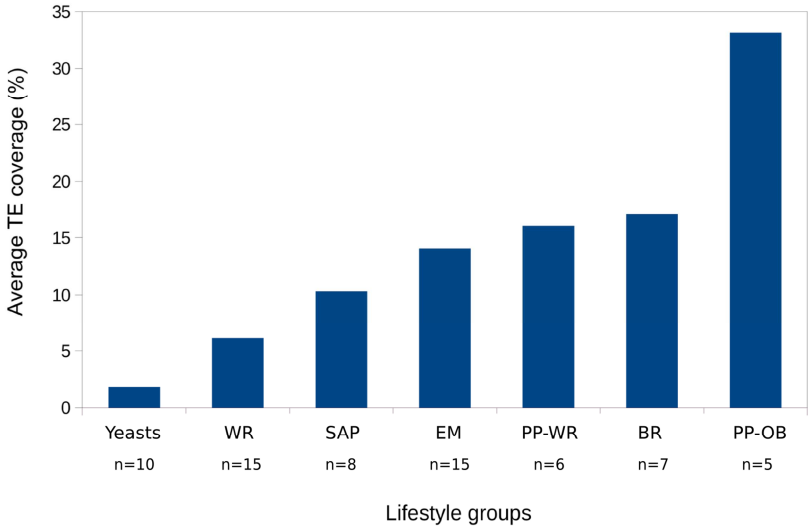

Figure 2. Average TE content of species-groups according to the primary basidiomycete lifestyle. n - Number of species in the group; WR – White rots; SAP - Saprotrophs (filamentous); EM – Ectomycorrhizal; PP-WR - Plant-pathogenic white-rots; BR – Brown rots; and PP-OB – plant pathogenic obligate biotrophs.

5.6. Dealing with unwanted repeats: genome defense

Despite occasional benefits that TEs can promote in their hosts, they represent an important source of instability, and basidiomycetes carry defense mechanisms aimed to avoid their expansion. The complete absence of expression by TE blocks described in *P. ostreatus* suggests that they may be isolated in areas of heterochromatin or silenced by epigenetic mechanisms (Castanera et al. 2016). In this sense, TE clusters of *L. bicolor* were shown to be transcriptionally repressed and highly methylated (Zemach et al. 2010). Cytosine methylation is known to occur in ascomycetes (Binz et al. 1998, Montanini et al. 2014; Jeon et al. 2015) and basidiomycetes (Binz et al. 1998, Foulongne-Oriol et al. 2013, Zemach et al. 2010). It occurs especially at CG sites, and is an efficient means to shut-down TE expression. This process, however, is not permanent; for example, a hypothetical loss of function mutation of the involved methyltransferase gene could lead to the recovery of TE activity. In ascomycetes, DNA methylation is often linked to RIP (repeat-induced point) mutations (Selker et al. 2003). This mechanism produces C to T mutations in repetitive elements during sexual reproduction and leads to their permanent inactivation. Horns et al. (2012) assessed RIP mutations in eight basidiomycetes from three subphyla and concluded that only TE sequences from members of

Pucciniomycotina have evidence of RIP mutations. In these species, which include the plant pathogens *P. graminis*, *M. larici-populina*, and *R. graminis* as well as the anther smut fungus *M. lychnidis-dioicae*, the authors found a hypermutation pattern that preferentially targets TpCpG sites. Interestingly, another study identified RIP mutations in *Ustilago hordei* (Laurie et al. 2012), despite RIP is absent in its closely related species *U. maydis*.

Finally, two main mechanisms of fungal RNA interference (RNAi) have been described to trigger post-transcriptional silencing of TEs in fungi: quelling and MSUD (meiotic silencing of unpaired DNA). Both methods are based on the detection of aberrant RNAs (such as those derived from transposons) and are dependent on Argonaute/Dicer components. MSUD occurs when chromosomal regions are unpaired during meiosis, whereas quelling triggers the silencing of homologous genes present in tandem arrays during the vegetative phase. The main components of RNAi machinery have been detected in several species distributed throughout the basidiomycete phylogeny, suggesting the presence of this pathway. A recent *in silico* study of orthologous RNAi genes in a dataset of 33 basidiomycete genomes did not find evidence for the presence of the MSUD pathway, although it concluded that quelling probably exists in basidiomycetes (Hu et al. 2013). One of the few studies examining RNAi in basidiomycetes described SIS (sexual-induced silencing) in *C. neoformans*, a mechanism dependent on Argonaute, Dicer, and a RNA-directed RNA polymerase that occurs primarily during sexual reproduction (Wang et al. 2010). This mechanism was shown to produce post-transcriptional silencing of repetitive transgenes and TEs, such as LTR-retrotransposons present in centromeres. Additionally, a mechanism similar to quelling was described in *C. neoformans* and named MIT (mitotic-induced silencing). In this case, high-copy, integrated transgenes led to RNAi-mediated silencing of homologous sequences during vegetative growth (Wang et al. 2012). In this regards, a recent study from our group has described the presence and transcriptional activity of the *core* proteins of the RNAi pathway in *P. ostreatus* (Borgognone et al. 2017). This finding suggests that *P. ostreatus* is probably able to epigenetically inactivate TEs. In addition, in the same study we describe the mobilization of elements from the HELPO2 family (described in Chapter II) in subclones of the strain PC15, representing the first evidence of somatic transposition of a native helitron on its living host.

5.7. Applications of TEs in basidiomycetes

Given the inherent potential of TEs to mobilize and integrate into new loci, they have been used as powerful genetic tools for identifying the function of unknown genes, a methodology commonly known as transposon tagging (Daboussi and Capy 2003). Transposons have been

widely used in eukaryotes to map and isolate targeted genes in natural and heterologous hosts. This allows for genotype-to-phenotype studies by cloning the gene in which the TE was inserted (Romao and Hamer 1992; Dioh et al. 2000). For fungi, *U. maydis* was used as a model system to study the first heterologous transposition in a basidiomycete; the authors used a *Tc1*/*mariner* transposon from *Caenorhabditis elegans* (Ladendorf et al. 2003). Later, mutagenesis mediated by *in vitro* transposition following biolistic transformation was reported for *C. neoformans* and *C. gattii*, two basidiomycete species that cause disease in immunocompromised individuals. This method can be used to disrupt genes efficiently and was described as a good alternative to *in vivo* transposition for knocking-out genes in basidiomycetes (Hu and Kronstad 2006). Fungal TEs used for transposon tagging are often cut-and-paste DNA transposons, such *Impala* (Hua-Van et al. 2001), *Fot1* (Migheli et al. 1999), or *Restless* (Kempken and Kück 1996). Such elements can excise and re-integrate into a new location making it possible to use two selectable markers. For example, one gene is inactivated by insertion of the transposon and it is restored upon excision; another selectable marker is used to identify the new insertions (Weld et al. 2006). Most studies utilizing fungal transposon-tagging have been performed in ascomycetes; however, the vast amount of information acquired recently from sequenced basidiomycete genomes offers a great opportunity to identify ideal candidates to be used as genetic tools. In this sense, screening TE annotations for highly active transposons would be a straightforward approach to identify potential candidates. Some *Pucciniomycotina* genomes could be an interesting starting point because they carry highly-expanded families of cut-and-paste TEs, as discussed above. Although functional annotation databases are increasing rapidly, the function remains unknown for approximately half of the annotated genes in basidiomycete genomes (Floudas et al. 2012), suggesting that every functional characterization of an unknown gene represents an important advancement for the field. In addition to transposon tagging, the ability of TEs to generate genome polymorphisms make them very useful as molecular markers. Several systems have been developed for fingerprinting, diversity analysis, or genetic linkage mapping based on the presence/absence of retrotransposons in a given locus, such as S-SAP (sequence-specific amplification polymorphism), IRAP (inter-retrotransposon amplified polymorphism), or REMAP (retrotransposon-microsatellite amplified polymorphism) (reviewed in (Kalendar 2011)). LTR-retrotransposons are the most abundant type of TE in most basidiomycetes. Normally, there are hundreds of these transposons per genome, although the distribution is highly variable between species (Muszewska et al. 2011). This makes such techniques interesting for most basidiomycete species. In this sense, the REMAP approach was proven to be a potent strategy in the taxonomic classification of higher fungal

groups. A *gypsy*-like *marY1* retroelement was used as a molecular marker for REMAP DNA fingerprinting that aided in the identification of 10 mushroom species. Primers used for this method generated specific patterns that allowed different species and strains of the same species to be distinguished from each other. Interestingly, this methodology enabled the differentiation of 14 commercially available cultivars of *P. ostreatus* and 16 of *P. eryngii* with high reproducibility, suggesting that these markers are a promising alternative to RAPD fingerprinting to identify edible mushrooms (Le et al. 2008). Genomic sequences of the most commercially-important mushroom species (*A. bisporus* and *P. ostreatus*) are already available, along with their corresponding TE annotations (Morin et al. 2012; Foulongne-Oriol et al. 2013; Riley et al. 2014; Castanera et al. 2016), which can greatly facilitate future advances in this direction. Studies on the ectomycorrhizal basidiomycete, *Tricholoma matsutake*, led to the discovery and characterization of the *marY1* LTR-retrotransposon. LTRs of this element were used as molecular markers for IRAP analysis to detect the genetic variability of this species (Murata et al. 2005). In this respect, it is worth mentioning that the use of such a LTR-based PCR system allowed for the detection of targeted species, while avoiding the generation of patterns that correspond to the plant host genome or other closely related fungi. In addition, modification of the IRAP method and use of internal regions (reverse-transcriptase) allowed for the identification and characterization of *Copia* LTR-retroelements in the ectomycorrhizal basidiomycetes *Pisolithus* spp. and *L. bicolor* (Díez et al. 2003). These examples show the suitability of using retrotransposon-based, species-specific markers to study symbiotic and pathogenic species whose DNA cannot be physically separated from that of their host.

5.8. References

Ali S, Laurie JD, Linning R, Cervantes-Chávez JA, Gaudet D, Bakkeren G (2014) An Immunity-Triggering Effector from the Barley Smut Fungus *Ustilago hordei* Resides in an Ustilaginaceae-Specific Cluster Bearing Signs of Transposable Element-Assisted Evolution. PLoS Pathog. doi: 10.1371/journal.ppat.1004223

Amselem J, Lebrun M-H, Quesneville H (2015) Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics 16:141. doi: 10.1186/s12864-015-1347-1

Bao D, Gong M, Zheng H, Chen M, Zhang L, Wang H, Jiang J, Wu L, Zhu Y, Zhu G, Zhou Y, Li C, Wang S, Zhao Y, Zhao G, Tan Q (2013) Sequencing and Comparative Analysis of the Straw Mushroom (*Volvariella volvacea*) Genome. PLoS One. doi: 10.1371/journal.pone.0058294

Binder M, Justo a, Riley R, Salamov a, Lopez-Giraldez F, Sjokvist E, Copeland a, Foster B, Sun H, Larsson E, Larsson K-H, Townsend J, Grigoriev I V, Hibbett DS (2013) Phylogenetic and phylogenomic overview of the Polyporales. Mycologia 105:1350–1373. doi: 10.3852/13-003

Binz T, D'Mello N, Horgen PA (1998) A Comparison of DNA Methylation Levels in Selected Isolates of Higher Fungi. Mycologia 90:785. doi: 10.2307/3761319

Borgognone A, Castanera R, Muguerza E, Pisabarro AG, Ramírez L (2017) Somatic transposition and meiotically driven elimination of an active helitron family in *Pleurotus ostreatus*. DNA Res dsw060. doi: 10.1093/dnares/dsw060

Branco S, Gladieux P, Ellison CE, Kuo A, Labutti K, Lipzen A, Grigoriev I V., Liao HL, Vilgalys R, Peay KG, Taylor JW, Bruns TD (2015) Genetic isolation between two recently diverged populations of a symbiotic fungus. Mol Ecol 24:2747–2758. doi: 10.1111/mec.13132

Canapa A, Barucca M, Biscotti MA, Forconi M, Olmo E (2016) Transposons, genome size, and evolutionary insights in animals. Cytogenet. Genome Res. 147:217–239. doi: 10.1159/000444429

Cantu D, Govindarajulu M, Kozik A, Wang M, Chen X, Kojima KK, Jurka J, Michelmore RW, Dubcovsky J (2011) Next generation sequencing provides rapid access to the genome of *Puccinia striiformis f.* sp. *tritici*, the causal agent of wheat stripe rust. PLoS One. doi: 10.1371/journal.pone.0024230

Castanera R, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, Grimwood J, Pérez G, Pisabarro AG, Grigoriev I V, Stajich JE, Ramírez L (2016) Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLoS Genet 12:e1006108. doi: 10.1371/journal.pgen.1006108

Castanera R, Pérez G, López L, Sancho R, Santoyo F, Alfaro M, Gabaldón T, Pisabarro AG, Oguiza JA, Ramírez L (2014) Highly expressed captured genes and cross-kingdom domains present in Helitrons create novel diversity in *Pleurotus ostreatus* and other fungi. BMC Genomics 15:1071. doi: 10.1186/1471-2164-15-1071

Chuong EB, Elde NC, Feschotte C (2016) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18:71–86. doi: 10.1038/nrg.2016.139

Chen S, Xu J, Liu C, Zhu Y, Nelson DR, Zhou S, Li C, Wang L, Guo X, Sun Y, Luo H, Li Y, Song J, Henrissat B, Levasseur A, Qian J, Li J, Luo X, Shi L, He L, Xiang L, Xu X, Niu Y, Li Q, Han M V, Yan H, Zhang J, Chen H, Lv A, Wang Z, Liu M, Schwartz DC, Sun C (2012) Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. Nat Commun 3:913. doi: 10.1038/ncomms1923

Chénais B, Caruso A, Hiard S, Casse, N (2012) The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. Gene. https://doi.org/10.1016/j.gene.2012.07.042

Daboussi M-J, Capy P (2003) Transposable elements in filamentous fungi. Annu Rev Microbiol 57:275–299. doi: 10.1146/annurev.micro.57.030502.091029

Díez J, Béguiristain T, Le Tacon F, Casacuberta JM, Tagu D, Jesu TB, Tacon L (2003) Identification of Ty1-copia retrotransposons in three ectomycorrhizal basidiomycetes : evolutionary relationships and use as molecular markers. Curr Genet 43:34–44. doi: 10.1007/s00294-002-0363-2

Dioh W, Tharreau D, Notteghem JL, Orbach M, Lebrun MH (2000) Mapping of avirulence genes in the rice blast fungus, *Magnaporthe grisea*, with RFLP and RAPD markers. Mol Plant Microbe Interact 13:217–227. doi: 10.1094/MPMI.2000.13.2.217

Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feaue N, Field M, Frey P, Gelhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kües U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van De Peer Y, Rouzé P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev I V, Szabo LJ, Martin F (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci U S A 108:9166–9171. doi: 10.1073/pnas.1019315108

Dutheil JY, Mannhaupt G, Schweizer G, Sieber CM, Münsterkötter M, Güldener U, Schirawski J, Kahmann R (2016) A tale of genome compartmentalization: the evolution of virulence clusters in smut fungi. Genome Biol Evol 8:681–704. doi: 10.1093/gbe/evw026

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18. doi: 10.1186/1471-2105-9-18

Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, Aerts A, Asiegbu FO, Baker SE, Barry K, Bendiksby M, Blumentritt M, Coutinho PM, Cullen D, De Vries RP, Gathman A, Goodell B, Henrissat B, Ihrmark K, Kauserud H, Kohler A, LaButti K, Lapidus A, Lavin JL, Lee Y-H, Lindquist E, Lilly W, Lucas S, Morin E, Murat C, Oguiza JA, Park J, Pisabarro AG, Riley R, Rosling A, Salamov A, Schmidt O, Schmutz J, Skrede I, Stenlid J, Wiebenga A, Xie X, Kües U, Hibbett DS, Hoffmeister D, Högberg N, Martin F, Grigoriev I V, Watkinson SC (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science (80-) 333:762–765. doi: 10.1126/science.1205411

Fernandez-Fueyo E, Ruiz-Dueñas FJ, Ferreira P, Floudas D, Hibbett DS, Canessa P, Larrondo LF, James TY, Seelenfreund D, Lobos S, Polanco R, Tello M, Honda Y, Watanabe T, Ryu JS, Kubicek CP, Schmoll M, Gaskell J, Hammel KE, St John FJ, Vanden Wymelenberg A, Sabat G, BonDurant SS, Syed K, Yadav JS, Doddapaneni H, Subramanian V, Lavín JL, Oguiza JA, Perez G, Pisabarro AG, Ramirez L, Santoyo F, Master E, Coutinho PM, Henrissat B, Lombard V, Magnuson JK, Kües U, Hori C, Igarashi K, Samejima M, Held BW, Barry KW, LaButti KM, Lapidus A, Lindquist EA, Lucas SM, Riley R, Salamov AA, Hoffmeister D, Schwenk D, Hadar Y, Yarden O, De Vries RP, Wiebenga A, Stenlid J, Eastwood D, Grigoriev I V, Berka RM, Blanchette RA, Kersten P, Martinez AT, Vicuna R, Cullen D (2012) Comparative genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis . Proc Natl Acad Sci U 109:5458-5463. doi: 10.1073/pnas. 1119912109

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: The protein families database. Nucleic Acids Res. 42: D222-30. doi: 10.1093/nar/gkt1223

Firrincieli A, Otillar R, Salamov A, Schmutz J, Khan Z, Redman RS, Fleck ND, Lindquist E,

Grigoriev I V, Doty SL (2015) Genome sequence of the plant growth promoting endophytic yeast *Rhodotorula graminis* WP1. Front Microbiol 6:978. doi: 10.3389/fmicb.2015.00978

Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martinez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Gorecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kues U, Kumar TKA, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Duenas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, St. John F, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev I V., Hibbett DS (2012) The Paleozoic Origin of Enzymatic Lignin Decomposition Reconstructed from 31 Fungal Genomes. Science (80-) 336:1715–1719. doi: 10.1126/science.1221748

Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One. doi: 10.1371/journal.pone.0016526

Foulongne-Oriol M, Murat C, Castanera R, Ramírez L, Sonnenberg ASM (2013) Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus*. Fungal Genet Biol 55:6–21. doi: 10.1016/j.fgb.2013.04.003

Gaskell J, Van den Wymelenberg A, Cullen D (1995) Structure, inheritance, and transcriptional effects of Pce1, an insertional element within *Phanerochaete chrysosporium* lignin peroxidase gene lipI. Proc Natl Acad Sci U S A 92:7465–7469. doi: 10.1073/pnas.92.16.7465

Hane JK, Anderson JP, Williams AH, Sperschneider J, Singh KB (2014) Genome Sequencing and Comparative Genomics of the Broad Host-Range Pathogen *Rhizoctonia solani* AG8. PLoS Genet. doi: 10.1371/journal.pgen.1004281

Hess J, Skrede I, Wolfe BE, Butti K La, Ohm RA, Grigoriev I V, Pringle A (2014) Transposable element dynamics among asymbiotic and ectomycorrhizal *Amanita* fungi. Genome Biol Evol 6:1564–1578. doi: 10.1093/gbe/evu121

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lücking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kõljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schüßler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N (2007) A higher-level phylogenetic classification of the Fungi. Mycol Res 111:509–547. doi: 10.1016/j.mycres.2007.03.004

Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H (2014) PASTEC: An automatic transposable element classification tool. PLoS One. doi: 10.1371/journal.pone.0091929

Hood ME (2005) Repetitive DNA in the automictic fungus *Microbotryum violaceum*. Genetica 124:1–10. doi: 10.1007/s10709-004-6615-y

Horns F, Petit E, Yockteng R, Hood ME (2012) Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi. Genome Biol Evol 4:240–247. doi: 10.1093/gbe/evs005

Hu G, Kronstad JW (2006) Gene disruption in *Cryptococcus neoformans* and *Cryptococcus gattii* by in vitro transposition. Curr Genet 49:341–350. doi: 10.1007/s00294-005-0054-x

Hu Y, Stenlid J, Elfstrand M, Olson A (2013) Evolution of RNA interference proteins dicer and argonaute in Basidiomycota. Mycologia 105:1489–98. doi: 10.3852/13-171

Hua-Van A, Pamphile JA, Langin T, Daboussi MJ (2001) Transposition of autonomous and engineered impala transposons in *Fusarium oxysporum* and a related species. Mol Gen Genet 264:724–731. doi: 10.1007/s004380000395

Janbon G, Ormerod KL, Paulet D, Byrnes EJ, Yadav V, Chatterjee G, Mullapudi N, Hon CC, Billmyre RB, Brunel F, Bahn YS, Chen W, Chen Y, Chow EWL, Coppée JY, Floyd-Averette A, Gaillardin C, Gerik KJ, Goldberg J, Gonzalez-Hilarion S, Gujja S, Hamlin JL, Hsueh YP, Ianiri G, Jones S, Kodira CD, Kozubowski L, Lam W, Marra M, Mesner LD, Mieczkowski PA, Moyrand F, Nielsen K, Proux C, Rossignol T, Schein JE, Sun S, Wollschlaeger C, Wood IA, Zeng Q, Neuvéglis C, Newlon CS, Perfect JR, Lodge JK, Idnurm A, Stajich JE, Kronstad JW, Sanyal K, Heitman J, Fraser JA, Cuomo CA, Dietrich FS (2014) Analysis of the Genome and Transcriptome of *Cryptococcus neoformans* var. *grubii* Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation. PLoS Genet. doi: 10.1371/journal.pgen.1004261

Jeon J, Choi J, Lee G-W, Park S-Y, Huh A, Dean R a, Lee Y-H (2015) Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, *Magnaporthe oryzae*. Sci Rep 5:8567. doi: 10.1038/srep08567

Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467. doi: 10.1159/000084979

Kalendar R (2011) The use of retrotransposon-based molecular markers to analyze genetic diversity. Ratar i Povrt 48:261–274. doi: 10.5937/ratpov1102261K

Kämper J, Kahmann R, Bölker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O, Perlin MH, Wösten HAB, de Vries R, Ruiz-Herrera J, Reynaga-Peña CG, Snetselaar K, McCann M, Pérez-Martín J, Feldbrügge M, Basse CW, Steinberg G, Ibeas JI, Holloman W, Guzman P, Farman M, Stajich JE, Sentandreu R, González-Prieto JM, Kennell JC, Molina L, Schirawski J, Mendoza-Mendoza A, Greilinger D, Münch K, Rössel N, Scherer M, Vraneš M, Ladendorf O, Vincon V, Fuchs U, Sandrock B, Meng S, Ho ECH, Cahill MJ, Boyce KJ, Klose J, Klosterman SJ, Deelstra HJ, Ortiz-Castellanos L, Li W, Sanchez-Alonso P, Schreier PH, Häuser-Hahn I, Vaupel M, Koopmann E, Friedrich G, Voss H, Schlüter T, Margolis J, Platt D, Swimmer C, Gnirke A, Chen F, Vysotskaia V, Mannhaupt G, Güldener U, Münsterkötter M, Haase D, Oesterheld M, Mewes HW, Mauceli EW, DeCaprio D, Wade CM, Butler J, Young S, Jaffe DB, Calvo S, Nusbaum C, Galagan J, Birren BW (2006) Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444 (7115):97–101. doi: 10.1038/nature05248

Kempken F, Kück U (1996) restless, an active Ac-like transposon from the fungus *Tolypocladium inflatum*: structure, expression, and alternative RNA splicing. Mol Cell Biol 16:6563–72. doi: 10.1128/MCB.16.11.6563

Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, Canbäck B, Choi C, Cichocki N, Clum A, Colpaert J, Copeland A, Costa MD, Doré J, Floudas D, Gay G, Girlanda M, Henrissat B, Herrmann S, Hess J, Högberg N, Johansson T, Khouja H-R, LaButti K, Lahrmann U, Levasseur A, Lindquist EA, Lipzen A, Marmeisse R, Martino E, Murat C, Ngan CY, Nehls U, Plett JM, Pringle A, Ohm RA, Perotto S, Peter M, Riley R, Rineau F, Ruytinx J, Salamov A, Shah F, Sun H, Tarkka M, Tritt A, Veneault-Fourrey C, Zuccaro A, Tunlid A, Grigoriev I V, Hibbett DS, Martin F (2015) Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet 47:410–5. doi: 10.1038/ng.3223

Konishi M, Hatada Y, Horiuchi J.-i (2013) Draft genome sequence of the Basidiomycetous yeast-Like Fungus *Pseudozyma hubeiensis* SY62, which produces an abundant amount of the biosurfactant mannosylerythritol lipids. Genome Announcements 1(4):e00409-13-e00409-13. doi:

10.1128/genomeA.00409-13

Kourist R, Bracharz F, Lorenzen J, Kracht ON, Chovatia M, Daum C, Deshpande S, Lipzen A, Nolan M, Ohm RA, Grigoriev I V., Sun S, Heitman J, Brück T, Nowrousian M (2015) Genomics and transcriptomics analyses of the oil-accumulating basidiomycete yeast *Trichosporon oleaginosus*: Insights into substrate utilization and alternative evolutionary trajectories of fungal mating systems. MBio. doi: 10.1128/mBio.00918-15

Labbe J, Murat C, Morin E, Tuskan GA, Le Tacon F, Martin F (2012) Characterization of transposable elements in the ectomycorrhizal fungus *Laccaria bicolor*. PLoS One 7:e40197. doi: 10.1371/journal.pone.0040197

Ladendorf O, Brachmann A, Kämper J (2003) Heterologous transposition in *Ustilago maydis*. Mol Genet Genomics 269:395–405. doi: 10.1007/s00438-003-0848-9

Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Güldener U, Münsterkötter M, Moore R, Kahmann R, Bakkeren G, Schirawski J (2012) Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. Plant Cell 24:1733–45. doi: 10.1105/tpc.112.097261

Le Q V, Won HK, Lee TS, Lee CY, Lee HS, Ro HS (2008) Retrotransposon microsatellite amplified polymorphism strain fingerprinting markers applicable to various mushroom species. Mycobiology 36:161–166. doi: 10.4489/MYCO.2008.36.3.161

Lee S-I, Kim N-S (2014) Transposable elements and genome size variations in plants. Genomics Inform 12:87–97. doi: 10.5808/GI.2014.12.3.87

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104:520–533. doi: 10.1038/hdy.2009.165

Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA, Allen JE, Bosdet IE, Brent MR, Chiu R, Doering TL, Dontin MJ, D'Souza CA, Fox DS, Grinberg V, Fu JM, Fukushima M, Haas BJ, Huang JC, Janbon G, Jones SJM, Koo HL, Krzywinski MI, Kwon-Chung JK, Lengeler KB, Maiti R (2005) The Genome of the Basidiomycetous Yeast and Human Pathogen *Cryptococcus neoformans*. Science (80-) 307:1321–1324. doi: 10.1126/science.1103773

Lynch M (2007) The Origins of Genome Architecture 2007. Science 302:1401–1404. doi: 10.1126/science.1089370

Lynch M, Conery JS (2003) The origins of genome complexity. Science (80-) 302:1401–1404. doi: 10.1126/science.1089370

Martin F, Aerts A, Ahrén D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buée M, Brokstein P, Canbäck B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbé J, Lin YC, Legué V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Oudot-Le Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kües U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouzé P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452(7183):88–92. doi: 10.1038/nature06556

Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J, Anthouard V, Arcioni S, Artiguenave F, Aury J-M, Ballario P, Bolchi A, Brenna A, Brun A, Buee M, Cantarel B, Chevalier G, Couloux A,

Da Silva C, Denoeud F, Duplessis S, Ghignone S, Hilselberger B, Iotti M, Marcais B, Mello A, Miranda M, Pacioni G, Quesneville H, Riccioni C, Ruotolo R, Splivallo R, Stocchi V, Tisserant E, Viscomi AR, Zambonelli A, Zampieri E, Henrissat B, Lebrun M-H, Paolocci F, Bonfante P, Ottonello S, Wincker P (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464:1033–1038. doi: 10.1038/nature08867

Martinez D, Larrondo LF, Putnam N, Gelpke MDS, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, Rokhsar D (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat Biotechnol 22:695–700. doi: 10.1038/nbt967

McCarthy EM, McDonald JF (2003) LTR STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367. doi: 10.1093/bioinformatics/btf878

Migheli Q, Laugé R, Davière JM, Gerlinger C, Kaper F, Langin T, Daboussi MJ (1999) Transposition of the autonomous Fot1 element in the filamentous fungus *Fusarium oxysporum*. Genetics 151:1005–1013. doi: 10.1093/nar/gkt1223

Min B, Park H, Jang Y, Kim JJ, Kim KH, Pangilinan J, Lipzen A, Riley R, Grigoriev I V., Spatafora JW, Choi IG (2015) Genome sequence of a white rot fungus *Schizopora paradoxa* KUC8140 for wood decay and mycoremediation. J Biotechnol 211:42–43. doi: 10.1016/j.jbiotec.2015.06.426

Montanini B, Chen P-Y, Morselli M, Jaroszewicz A, Lopez D, Martin F, Ottonello S, Pellegrini M (2014) Non-exhaustive DNA methylation-mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. Genome Biol 15:411. doi: 10.1186/s13059-014-0411-5

Morin E, Kohler A, Baker AR, Foulongne-Oriol M, Lombard V, Nagy LG, Ohm RA, Patyshakuliyeva A, Brun A, Aerts AL, Bailey AM, Billette C, Coutinho PM, Deakin G, Doddapaneni H, Floudas D, Grimwood J, Hildén K, Kües U, Labutti KM, Lapidus A, Lindquist EA, Lucas SM, Murat C, Riley RW, Salamov AA, Schmutz J, Subramanian V, Wösten HAB, Xu J, Eastwood DC, Foster GD, Sonnenberg ASM, Cullen D, de Vries RP, Lundell T, Hibbett DS, Henrissat B, Burton KS, Kerrigan RW, Challen MP, Grigoriev I V, Martin F (2012) Genome sequence of the button mushroom *Agaricus bisporus* reveals mechanisms governing adaptation to a humic-rich ecological niche. Proc Natl Acad Sci U S A 109:17501–6. doi: 10.1073/pnas.1206847109

Morita T, Koike H, Koyama Y, Hagiwara H, Ito E, Fukuoka T, Imura T, Machida M, Kitamoto D (2013) Genome sequence of the Basidiomycetous yeast *Pseudozyma antarctica* T-34, a producer of the glycolipid biosurfactants mannosylerythritol lipids. Genome Announcements 1(2):e00064-13-e00064-13. doi: 10.1128/genomeA.00064-13

Murata H, Babasaki K, Yamada A (2005) Highly polymorphic DNA markers to specify strains of the ectomycorrhizal basidiomycete *Tricholoma matsutake* based on sigmamarY1, the long terminal repeat of gypsy-type retroelement marY1. Mycorrhiza 15:179–186. doi: 10.1007/s00572-004-0319-0

Murata H, Miyazaki Y, Babasaki K (2001) The Long Terminal Repeat (LTR) sequence of marY1, a retroelement from the ectomycorrhizal homobasidiomycete *Tricholoma matsutake*, is highly conserved in various higher fungi. Bioscience, Biotechnology, and Biochemistry 65(10):2297-2300. doi: 10.1271/bbb.65.2297

Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR retrotransposons in fungi. PLoS One 6:e29425. doi: 10.1371/journal.pone.0029425

Muszewska A, Steczkiewicz K, Ginalski K (2013) DIRS and Ngaro Retrotransposons in Fungi. PLoS One. doi: 10.1371/journal.pone.0076319

Nemri A, Saunders DGO, Anderson C, Upadhyaya NM, Win J, Lawrence GJ, Jones DA, Kamoun S, Ellis JG, Dodds PN (2014) The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. Front Plant Sci 5:98. doi: 10.3389/fpls.2014.00098

Ohm RA, De Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, De Vries RP, Record E, Levasseur A, Baker SE, Bartholomew KA, Coutinho PM, Erdmann S, Fowler TJ, Gathman AC, Lombard V, Henrissat B, Knabe N, Kües U, Lilly WW, Lindquist E, Lucas S, Magnuson JK, Piumi F, Raudaskoski M, Salamov A, Schmutz J, Schwarze FWMR, Vankuyk PA, Horton JS, Grigoriev I V, Wösten HAB (2010) Genome sequence of the model mushroom *Schizophyllum commune*. Nat Biotechnol 28:957–963. doi: 10.1038/nbt.1643

Olson Å, Aerts A, Asiegbu F, Belbahri L, Bouzid O, Broberg A, Canbäck B, Coutinho PM, Cullen D, Dalman K, Deflorio G, van Diepen LTA, Dunand C, Duplessis S, Durling M, Gonthier P, Grimwood J, Fossdal CG, Hansson D, Henrissat B, Hietala A, Himmelstrand K, Hoffmeister D, Högberg N, James TY, Karlsson M, Kohler A, Kües U, Lee Y-H, Lin Y-C, Lind M, Lindquist E, Lombard V, Lucas S, Lundén K, Morin E, Murat C, Park J, Raffaello T, Rouzé P, Salamov A, Schmutz J, Solheim H, Ståhlberg J, Vélëz H, de Vries RP, Wiebenga A, Woodward S, Yakovlev I, Garbelotto M, Martin F, Grigoriev I V, Stenlid J (2012) Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen. New Phytol 194:1001–1013. doi: 10.1111/j.1469-8137.2012.04128.x

Padamsee M, Kumar TKA, Riley R, Binder M, Boyd A, Calvo AM, Furukawa K, Hesse C, Hohmann S, James TY, LaButti K, Lapidus A, Lindquist E, Lucas S, Miller K, Shantappa S, Grigoriev I V, Hibbett DS, McLaughlin DJ, Spatafora JW, Aime MC (2012) The genome of the xerotolerant mold *Wallemia sebi* reveals adaptations to osmotic stress and suggests cryptic sexual reproduction. Fungal Genet Biol 49:217–226. doi: 10.1016/j.fgb.2012.01.007

Perlin MH, Amselem J, Fontanillas E, Toh SS, Chen Z, Goldberg J, Duplessis S, Henrissat B, Young S, Zeng Q, Aguileta G, Petit E, Badouin H, Andrews J, Razeeq D, Gabaldón T, Quesneville H, Giraud T, Hood ME, Schultz DJ, Cuomo C a (2015) Sex and parasites: genomic and transcriptomic analysis of *Microbotryum lychnidis-dioicae*, the biotrophic and plant-castrating anther smut fungus. BMC Genomics 16:461. doi: 10.1186/s12864-015-1660-8

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:I351–I358. doi: 10.1093/bioinformatics/bti1018

Raffaele S, Kamoun S (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 10:417–430. doi: 10.1038/nrmicro2790

Riley R, Salamov AA, Brown DW, Nagy LG, Floudas D, Held BW, Levasseur A, Lombard V, Morin E, Otillar R, Lindquist EA, Sun H, LaButti KM, Schmutz J, Jabbour D, Luo H, Baker SE, Pisabarro AG, Walton JD, Blanchette RA, Henrissat B, Martin F, Cullen D, Hibbett DS, Grigoriev I V (2014) Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. Proc Natl Acad Sci U S A 111:9923–9928. doi: 10.1073/pnas.1400592111

Romao J, Hamer JE (1992) Genetic organization of a repeated DNA sequence family in the rice blast fungus. Proc Natl Acad Sci U S A 89:5316–20. doi: 10.1073/pnas.89.12.5316

Schirawski J, Mannhaupt G, Munch K, Brefort T, Schipper K, Doehlemann G, Di Stasio M, Rossel N, Mendoza-Mendoza A, Pester D, Muller O, Winterberg B, Meyer E, Ghareeb H, Wollenberg T, Munsterkotter M, Wong P, Walter M, Stukenbrock E, Guldener U, Kahmann R (2010) Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330(6010):1546–1548. doi: 10.1126/science.1195330

Selker EU, Tountas N a, Cross SH, Margolin BS, Murphy JG, Bird AP, Freitag M (2003) The methylated component of the *Neurospora crassa* genome. Nature 422:893–897. doi:

10.1038/nature01564

Sonnenberg ASM, Gao W, Lavrijssen B, Hendrickx P, Sedaghat-Tellgerd N, Foulongne-Oriol M, Kong W-S, Schijlen EGWM, Baars JJP, Visser RGF (2016) A detailed analysis of the recombination landscape of the button mushroom *Agaricus bisporus* var. *bisporus*. Fungal Genet Biol 93:35–45. doi: 10.1016/j.fgb.2016.06.001

Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Ver Loren van Themaat E, Brown JKM, Butcher SA, Gurr SJ, Lebrun M-H, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler L V, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, Lopez-Ruiz FJ, Lu X, Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, Sacristan S, Schmidt SM, Schon M, Skamnioti P, Sommer H, Stephens A, Takahara H, Thordal-Christensen H, Vigouroux M, Wessling R, Wicker T, Panstruga R (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science 330:1543–1546. doi: 10.1126/science.1194573

Stajich JE, Wilke SK, Ahrén D, Hang C, Birren BW, Borodovsky M, Burns C, James TY, Kamada T, Kilaru S, Kodira C, Kües U, Kupfer D, Kwan HS (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). Pnas 107:11889–11894. doi: 10.1073/pnas.1003391107

Suzuki H, MacDonald J, Syed K, Salamov A, Hori C, Aerts A, Henrissat B, Wiebenga A, vanKuyk PA, Barry K, Lindquist E, LaButti K, Lapidus A, Lucas S, Coutinho P, Gong Y, Samejima M, Mahadevan R, Abou-Zaid M, de Vries RP, Igarashi K, Yadav JS, Grigoriev I V, Master ER (2012) Comparative genomics of the white-rot fungi, *Phanerochaete carnosa* and *P. chrysosporium*, to elucidate the genetic basis of the distinct wood types they colonize. BMC Genomics 13:444. doi: 10.1186/1471-2164-13-444

Toome M, Ohm RA, Riley RW, James TY, Lazarus KL, Henrissat B, Albu S, Boyd A, Chow J, Clum A, Heller G, Lipzen A, Nolan M, Sandor L, Zvenigorodsky N, Grigoriev I V., Spatafora JW, Aime MC (2014) Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. New Phytol 202:554–564. doi: 10.1111/nph.12653

Zajc J, Liu Y, Dai W, Yang Z, Hu J, Gostinčar C, Gunde-Cimerman N (2013) Genome and transcriptome sequencing of the halophilic fungus *Wallemia ichthyophaga*: haloadaptations present and absent. BMC Genomics 14:617. doi: 10.1186/1471-2164-14-617

Wang X, Hsueh YP, Li W, Floyd A, Skalsky R, Heitman J (2010) Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi. Genes Dev 24:2566–2582. doi: 10.1101/gad.1970910

Wang X, Wang P, Sun S, Darwiche S, Idnurm A, Heitman J (2012) Transgene Induced Co-Suppression during Vegetative Growth in *Cryptococcus neoformans*. PLoS Genet. doi: 10.1371/journal.pgen.1002885

Weld RJ, Plummer KM, Carpenter M a, Ridgway HJ (2006) Approaches to functional genomics in filamentous fungi. Cell Res 16:31–44. doi: 10.1038/sj.cr.7310006

Wicker T, Oberhaensli S, Parlange F, Buchmann JP, Shatalina M, Roffler S, Ben-David R, Dolezel J, Simkova H, Schulze-Lefert P, Spanu PD, Bruggmann R, Amselem J, Quesneville H, Ver Loren van Themaat E, Paape T, Shimizu KK, Keller B (2013) The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat Genet 45:1092–1096. doi: 10.1038/ng.2704

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante

M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi: 10.1038/nrg2165

Wong LH, Choo KHA (2004) Evolutionary dynamics of transposable elements at the centromere. Trends Genet. 20:611–616. doi: 10.1016/j.tig.2004.09.011.

Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, DeAngelis YM, Reeder NL, Johnstone KR, Leland M, Fieno AM, Begley WM, Sun Y, Lacey MP, Chaudhary T, Keough T, Chu L, Sears R, Yuan B, Dawson TL (2007) Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. Proc Natl Acad Sci 104(47):18730–18735. doi: 10.1073/pnas.0706756104

Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916–9. doi: 10.1126/science.1186366

Zuccaro A, Lahrmann U, Güldener U, Langen G, Pfiffi S, Biedenkopf D, Wong P, Samans B, Grimm C, Basiewicz M, Murat C, Martin F, Kogel KH (2011) Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. PLoS Pathog. doi: 10.1371/journal.ppat.1002290

Chapter VI: Supplementary information

6.1. Chapter II:

Distribution, activity and functional characterization of helitron transposons in Pleurotus ostreatus

Supplementary methods

Phylogenetic reconstructions.

Sequence analyses were performed using MEGA5 (Kumar et al. 2008). Alignments were performed with Clustal Omega (Sievers et al. 2011). Phylogenetic analysis shown in Fig S1 was performed on JGI filtered-protein sequences using the Phylogeny.fr platform (Dereeper et al. 2008). Ambiguous regions were removed from multiple sequence alignment with Gblocks v0.91b (Talavera and Castresana 2007). A Phylogenetic tree was reconstructed using the maximum likelihood method implemented in the PhyML program v3.0 aLRT (Guindon et al. 2010).

Table S1. Primers used for RT-qPCR expression analyses.

Name	Sequence	Efficiency	Specificity *
HELPO1.1rep.fw	CCAGATGCCGAGATCAAGCTTCG	1.87	HELPO1 helicases (4)
HELPO1.1 rep.RV	GGCATAATCATGGCAACCTC		
HELPO1.2rep.fw	CCACCAACTCCCACAGAAT	1.86	HELPO1.2 helicases (2)
HELPO1.2rep.fw	TCTAGCACCCCCGATTTATG		
HELPO2_dg.fw	AAACTGCGGACTCCTGAAGA	1.87	HELPO2 helicases (6)
HELPO2_dg.rv	CAGCTGTGGTGCTTCCAGTA		
capA.fw	AAATGGACCCCTCCGTTTAC	1.98	capA genes (3) - HELPO1
capA.rv	TTTCTGCAAGGGACCCATAG		
capA2.fw	CCTGTTGTTGCATGATCCAG	1.89	capA2 gene (1) - HELPO1
capA2.rv	GATGTGCGCCTCAGTAGACA		
capB.fw	GGGCTTGCTGTATTGGAAAA	1.91	capB genes (2) - HELPO1.2
capB.rv	TGGGGAGCGAGATAGAATTG		
capC.fw	CACGAGCAATTTTTGCAATG	1.87	capC genes (6) - HELPO1.3
capC.rv	GTAAGGGTCCTGAGCAGCAG		
capD.fw	GCAGAGCAGCGAGAGTTTCT	/	capD gene (1) - HELPO1.3
capD.rv	AAAATCCCGGTACGTGTTCA		
capF.fw	CATTGGACTGGGAATCTGCT	/	capF gene (1) - HELPO1.3
capF.rv	CCCTGCTTTTTGACTTCAGC		
pep_fw	CTATCTCGGGAACGGTATATCA	1.89	PC15 V2.0 ID:1092697
pep_rv	CCGCTGGTACTGGTACTATAA		

* - In parenthesis is shown the number of gene copies in PC15 genome
/ - Efficiency could not be determined by linear regression due to the lack of enough sample

Table S2. Summary of helitron-like 3'- terminal ends found by HelSearch

Hairpin family	Location PC9 (*)	Location PC15 (*)
ACATTCGA_TG_TCGAATGT	scaffold_002:1177418-1181018 (+)	scaffold_04:2285016-2288616 (+)
	scaffold_002:1191260-1194860 (-)	scaffold_04:2297555-2301098 (-)
AGTGAT_GA_ATCACT	scaffold_002:196554-200154 (+)	scaffold_04:3240741-3244341 (+)
	scaffold_002:232065-235665 (-)	scaffold_04:3286520-3290120 (-)
CCCGTGC_TTC_GCACGGG	scaffold_091:1-659 (+)	scaffold_07:1391507-1395107 (+)
	scaffold_006:1348145-1351745 (-)	scaffold_07:1521433-1525033 (-)
	scaffold_004:2670410-2674010 (-)	
CCGTGC_GTA_GCACGG	scaffold_142:1-633 (+)	scaffold_01:1414547-1418047 (+)
	scaffold_478:273-3873 (-)	scaffold_01:3751228-3754828 (+)
	scaffold_007:1079010-1082610 (+)	scaffold_01:4537480-4541080 (-)
	scaffold_044:2511-6111 (-)	scaffold_02:1934771-1938371 (-)
		scaffold_05:1418338-1421838 (-)
		scaffold_07:378958-382559 (+)
		scaffold_07:2179834-2183434 (+)
		scaffold_08:26455-30055 (-)
		scaffold_08:423543-427143 (-)
		scaffold_11:760623-764223 (+)
TCTTAG_CC_CTAAGA	scaffold_006:1916618-1920218 (+)	scaffold_07:3019518-3023118(+)
	scaffold_006:1959378-1962978(-)	scaffold_07:3057094-3060694 (-)
AACCAG_GGC_CTGGTT		scaffold_06:103948-107548 (-)
		scaffold_06:167219-170819 (+)
AAGTGTG_CA_CACACTT		scaffold_06:1808078-1809368 (+)
		scaffold_08:1123002-1126602 (+)
CGTAGCCAC_ACT_GTGGCTACG		scaffold_02:2341430-2345030 (-)
		scaffold_02:2382940-2385356 (-)
		scaffold_02:2383133-2385356 (-)
		scaffold_02:2383195-2385356 (-)
CTTGTC_GA_GACAAG		scaffold_04: 387277-390877 (+)
		scaffold_04:591329-594929 (-)
GGGCAT_CG_ATGCCC		scaffold_06:350706-354306(-)
		scaffold_06:2250480-2254080 (-)
TCAGGG_CTT_CCCTGA		scaffold_02:1198298-1201899 (-)
		scaffold_07:41882-45482 (+)
AAATGC_CG_GCATTT	scaffold_212:206-3806 (+)	
	scaffold_003:2776155-2779755 (+)	
GGACGG_TAG_CCGTCC	scaffold_105:1-3039 (+)	
	scaffold_23:289293-292893 (-)	
TGTGGA_TGAG_TCCACA	scaffold_538:1-401 (+)	
	scaffold_007:1677183-1680783 (-)	
TTGCTC_ATC_GAGCAA	scaffold_011:11207-14807 (-)	
	scaffold_011:11359-14959 (-)	

Table S3. Matrix of nucleotide similarity between intact copies of HELPO1 family captured genes.

	capA	**capA**	**capD**	**capC**	**capE**	**capB**	**capB**	**capF**	**capA2**
capA	100	99.53	44.8	53.6	54.57	43.49	43.49	38.07	45.28
capA	99.53	100	44.83	53.36	54.72	43.43	43.43	37.91	45.33
capD	44.8	44.83	100	74.91	54.01	40.12	40.02	30.8	41.02
capC	53.6	53.36	74.91	100	62.16	46.08	46.37	41.71	49.33
capE	54.57	54.72	54.01	62.16	100	42.35	42.18	41.88	46.48
capB	43.49	43.43	40.12	46.08	42.35	100	99.23	47.54	60.84
capB	43.49	43.43	40.02	46.37	42.18	99.23	100	47.54	60.73
capF	38.07	37.91	30.8	41.71	41.88	47.54	47.54	100	74.98
capA2	45.28	45.33	41.02	49.33	46.48	60.84	60.73	74.98	100

Table S4. Description of conserved domains found in *P. ostreatus* helitrons

Helitron family	Name	ID	start	end	e value	Reading frame	Description
HELPO1.1	Helitron_like_N	pfam14214	2289	2838	1.81e-65	(+2)	Helitron helicase-like domain at N-terminus. Found in helitron eukaryotic transposons
	PIF1	pfam05970	4188	5235	3.69e-113	(+2)	PIF1-like helicase/This family includes a large number of largely uncharacterized plant proteins
	UvrD_C_2 super family	cl19402	5295	5514	5.38e-03	(+2)	UvrD-like helicase C-terminal domain; This domain is found at the C-terminus of a wide variety of helicase enzymes
HELPO1.2	PHA03255	PHA03255	636	1032	7.93e-03	(+2)	Provisional domain
	Pilt super family	pfam15453	636	1176	1.78e-04	(+3)	Pilt is a family of eukaryotic tight junction-proteins that binds to guanylate-kinase
	Helitron_like_N	pfam14214	2280	2829	2.29e-81	(+1)	Helitron helicase-like domain at N-terminus. Found in helitron eukaryotic transposons
	PIF1	pfam05970	4287	5334	3.67e-102	(+3)	PIF1-like helicase/This family includes a large number of largely uncharacterized plant proteins
	UvrD_C_2 super family	cl19402	5199	5559	2.14e-04	(+3)	UvrD-like helicase C-terminal domain; This domain is found at the C-terminus of a wide variety of helicase enzymes
HELPO1.2	Pilt super family	pfam15453	636	1176	3.70e-04	(+3)	Pilt is a family of eukaryotic tight junction-proteins that binds to guanylate-kinase
	Helitron_like_N	pfam14214	2280	2829	2.66e-82	(+1)	Helitron helicase-like domain at N-terminus. Found in helitron eukaryotic transposons
	CHROMO	cd00024	5514	5361	4.19e-09	(-2)	Chromatin organization modifier
	rve	pfam00665	6420	6069	2.62e-15	(-2)	Integrase core domain
	RNase_HI_RT_Ty3	cd09274	7281	6915	5.60e-51	(-2)	Ty3/Gypsy family of RNase HI in long-term repeat retroelements
	RVT_1	pfam00078	8040	7560	3.46e-29	(-2)	Reverse transcriptase (RNA-dependent DNA polymerase)
	RT_LTR	cd01647	8085	7560	1.33e-68	(-2)	Reverse transcriptases (RTs) from retrotransposons and retroviruses
	retropepsin_like	cd00303	8952	8697	1.15e-12	(-2)	Pepsin-like aspartate proteases
	zf-CCHC	pfam00098	9417	9375	7.08e-04	(-3)	Zinc knuckle binding motif
	Retrotrans_gag super family	pfam03732	10041	9762	7.99e-07	(-3)	Retrotransposon gag protein; Gag or Capsid-like proteins from LTR retrotransposons
	PHA03247	PHA03247	10491	9384	5.19e-13	(-1)	Large tegument protein UL36; Provisional
	Tymo_45kd_70kd	pfam03251	10692	9375	1.07e-06	(-2)	Tymovirus 45/70Kd protein; Tymoviruses are single stranded RNA viruses
	PIF1	pfam05970	11127	11883	3.21e-63	(+2)	PIF1-like helicase/This family includes a large number of largely uncharacterized plant proteins
	UvrD_C_2 super family	cl19402	11808	12168	2.81e-04	(+2)	UvrD-like helicase C-terminal domain; This domain is found at the C-terminus of a wide variety of helicase enzymes
helpo1.3	PHA03247	PHA03247	1380	1905	8.94e-04	(-2)	Large tegument protein UL36; Provisional

	Cauli_VI super family	pfam01693	5793	5838	2.52e-03	(+1)	Main component of viral inclusion bodies or viroplasms
	COG4328 super family	COG4328	7008	7146	0.01	(+3)	Predicted transposase
HELPO2	Mucin super family	pfam01456	1866	2013	5.17e-03	(+1)	Mucin-like glycoprotein; This family of trypanosomal proteins resemble vertebrate mucins
	Keratin_B2	pfam01500	1629	2007	8.47e-06	(+1)	Keratin, high sulfur B2 protein
	HpaP super family	cl17849	1758	2049	1.98e-04	(+2)	Type III secretion protein (HpaP)
	Membrane-FADS-like super family	cl00615	1851	1998	2.62e-03	(+3)	membrane fatty acid desaturase
	TT_ORF1 super family	pfam02956	1736	1973	1.11e-03	(-2)	TT viral orf 1; TT virus (TTV)
	AdoMet_MTases super family	cl17173	1883	1994	9.38e-05	(-3)	S-adenosylmethionine-dependent methyltransferases
	GP38 super family	pfam05268	1643	2003	1.48e-03	(-3)	Phage tail fibre adhesin Gp38
	Helitron_like_N	pfam14214	2979	3528	1.68e-63	(+2)	Helitron helicase-like domain at N-terminus. Found in helitron eukaryotic transposons
	PIF1	pfam05970	4878	5922	8.76e-124	(+2)	PIF1-like helicase/This family includes a large number of largely uncharacterized plant proteins
	AAA_30 super family	pfam13604	4875	5331	6.47e-05	(+2)	AAA domain; This family of domains contain a P-loop motif
	UvrD_C_2 super family	cl19402	5838	6180	7.03e-04	(+2)	UvrD-like helicase C-terminal domain; This domain is found at the C-terminus of a wide variety of helicase enzymes

Table S5. Matrix of nucleotide similarity between intact copies of the autonomous elements of the HELPO1 and HELPO2 families in PC15 genome.

	helpo 1.3	**helpo 1.3**	**helpo 1.3**	**helpo 1.3**	**helpo 1.3**	**helpo 1.3**	**helpo 1.3**	**HEL PO2**	**HEL PO2**	**HEL PO2**	**HEL PO2**	**HEL PO2**	**HELP O1.1**	**HELP O1.1**	**HELP O1.2**
helpo1.3	100	96.58	96.27	96.01	96.63	96.81	96.67	47.45	47.56	47.56	47.56	47.56	40.72	37.84	37.85
helpo1.3		100	98.63	98.5	99.06	99.02	98.89	46.94	47.05	47.05	47.05	47.05	41.21	39.33	38.42
helpo1.3			100	98.19	98.59	98.93	98.7	47.11	47.22	47.22	47.22	47.22	40.55	39.24	39.34
helpo1.3				100	98.46	98.58	98.45	47.01	47.11	47.11	47.11	47.11	40.94	39.15	38.59
helpo1.3					100	98.93	98.86	47.23	47.33	47.33	47.33	47.33	40.74	39	38.51
helpo1.3						100	99.87	47.23	47.33	47.33	47.33	47.33	40.87	39.32	38.51
helpo1.3							100	47.23	47.33	47.33	47.33	47.33	40.87	39.24	38.51
HELPO2								100	100	100	100	100	50.2	50.18	49.01
HELPO2									100	100	100	100	50.19	50.19	49.00
HELPO2										100	100	100	50.19	50.19	49.00
HELPO2											100	100	50.19	50.19	49.00
HELPO2												100	50.19	50.19	49.00
HELPO1.1													100	56.93	56.69
HELPO1.1														100	60.05
HELPO1.2															100

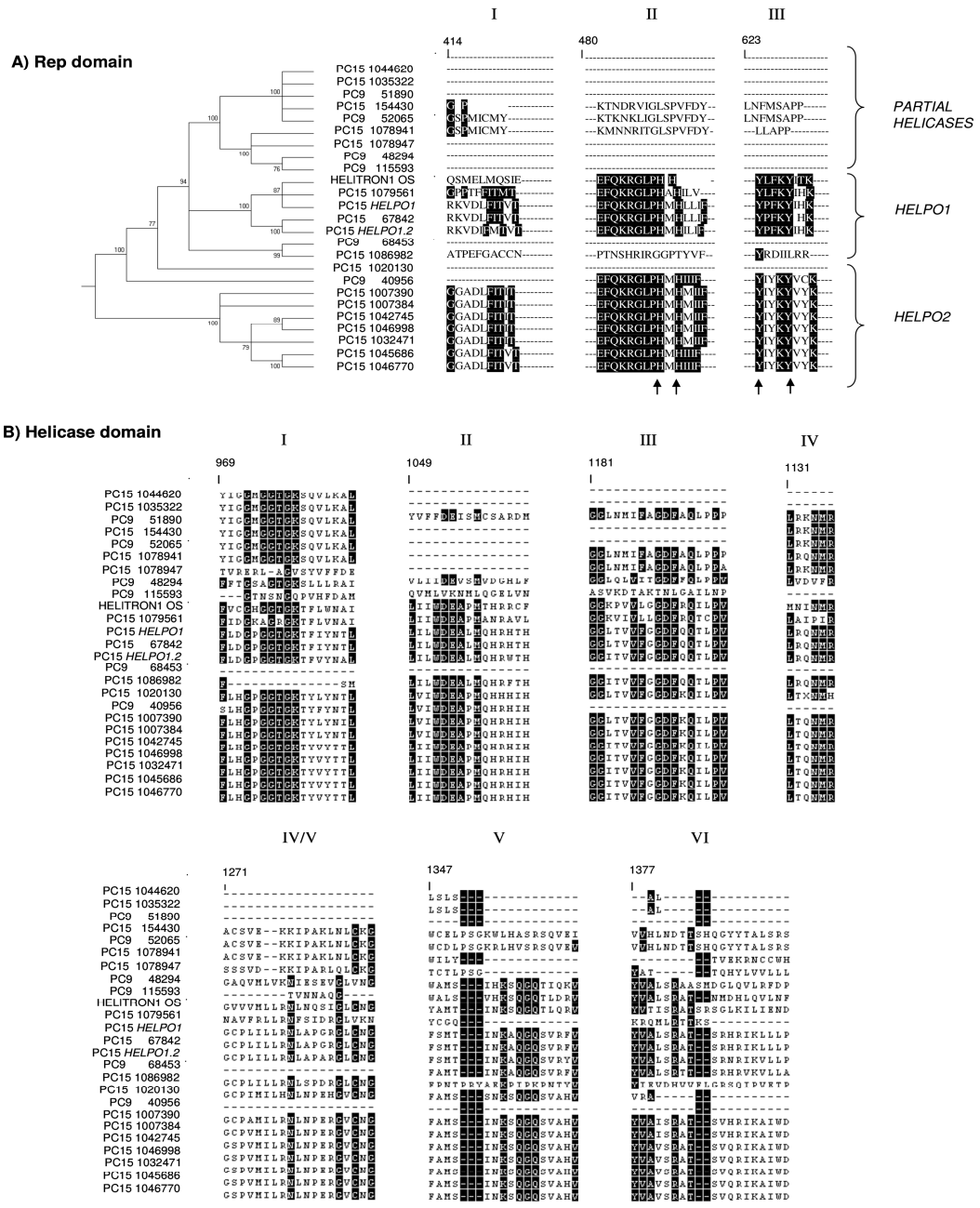

Figure S1. Predicted functional domains of *P.ostreatus* RepHel proteins. The conserved motifs of the Rep catalytic core described by (Kapitonov and Jurka 2007) are shown in black (A). The seven domains of the SF1 helicase superfamily found in helitrons of other species (Pritham and Feschotte 2007) are shown in B. Black represents more than 60 % similarity. The unrooted phylogenetic tree was constructed using MUSCLE and PhyML. HELITRON1 OS = *Oryza sativa* helitron.

References

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. doi: 10.1093/nar/gkn180

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59:307–321. doi: 10.1093/sysbio/syq010

Kapitonov V V, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23:521–529. doi: 10.1016/j.tig.2007.08.004

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299–306. doi: 10.1093/bib/bbn017

Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci 104:1895–1900. doi: 10.1073/pnas.0609601104

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi: 10.1038/msb.2011.75

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–77. doi: 10.1080/10635150701472164

6.2. Chapter III:

Transposable elements versus the fungal genome: impact on whole-genome architecture and transcriptional profiles

Supplementary methods

Phylogeny of LTR-retrotransposons conserved domains

Reverse transcriptase and RNAse domains of Gypsy and Copia elements present were extracted from LTR-retrotransposons of the TE library using exonerate (Slater and Birney 2005) and aligned with MUSCLE (Edgar 2004). The alignments were trimmed using trimAl (Capella-Gutierrez et al. 2009) with the default parameters, and an approximate maximum likelihood tree was constructed using FastTree (Price et al. 2009) and edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Expression and phylogeny of *P. ostreatus* DNA methyltransferases

Searches were performed in the PC15 and PC9 homepages of the JGI database for retrieving every protein classified under the GO term “DNA methylation”. A protein domain analysis was performed using the Conserved Domain Database (Marchler-Bauer and Bryant 2004), and only those carrying the Dcm domain (Site-specific DNA-cytosine methylase, COG0270) were retained. The Dim-2 DNA methyltransferase of Neurospora crassa (gi_28921348) was obtained from the NCBI database. The phylogenetic analysis was carried out using the protein sequences and the same methodology as for LTR-retrotransposons.

Table S1. Primers used for PCR amplification of polymorphic TE insertions.

	Forward	Reverse
Locus I	CGACTCCTCGGTGTCTGATT	ATACCCCAACGACAGTTTGC
Locus II	TCCTTTTCGCTGTCTTCCAT	GCACAGGGTCCCTAATCAAA
Locus III	AGAAGCAGCTGCCTGTCAAC	TTTTCTTGCTGTTCCGCTTT
Locus IV	CGCATGGTCGATGTCAATAA	CGGGTGCCTACGTGTTAAGT
Locus V	CGACAGCAGTTGCTGGAGTA	TGGCGGTAATAACCAAGGAG
Locus VI	TGACGGATTAGTTTCGAGCA	AGGCGTCTGTACCCGATCTA
Locus VII	TAAGGGTTTGGACCAAGCTG	CAAGCCCCATTTCATATGCT
Locus VIII	ATGTTACCTCCGTTGCCTTG	AAGACTGCGGTAGGCATTGT
Locus VIII	ATGTTACCTCCGTTGCCTTG	AAGACTGCGGTAGGCATTGT

Table S2. Differential Expression of orthologous genes displaying polymorphic TE insertions

TE insertions in PC15					
PC9 (noTE)	**PC15 (TE)**	**TE family**	**PC9 FPKM**	**PC15 FPKM**	**Description**
120845	1099919	Copia_2	13.1	0.14	Unknown function
100052	1106124	Gypsy_3	32.6	0.49	Unknown function
86999	160984	Copia_11	13.9	0.3	Unknown function
98979	154062	Gypsy_24	61.6	1.44	Unknown function
49583	166872	Gypsy_16	46.6	1.3	Unknown function
90025	158900	Gypsy_9	19.7	0.57	Unknown function
68169	1091908	Gypsy_23	1.5	0.12	Zinc finger, C2H2-type
91331	1044593	Gypsy_3	1.5	0.16	Glycoside Hydrolase Family 131 protein
89531	1097443	Copia_5	142.2	15.73	Unknown function
TE insertions in PC9					
PC9 (TE)	**PC15 (noTE)**	**TE family**	**PC9 FPKM**	**PC15 FPKM**	**Description**
95253	167769	Gypsy_3	0.3	26.39	Unknown function
91123	154703	Copia_7	8.4	147.56	GMC oxidoreductase
58056	1090089	HELPO1	9.4	94.15	Pyridoxamine 5'-phosphate oxidase
81558	152288	Copia_8	6.8	56.23	Dimeric alpha-beta barrel
117290	176657	Gypsy_19	11.4	89.85	Zinc finger, C2H2-type
125737	1085502	Gypsy_3	0.5	3.77	Cytochrome P450
126274	161195	DIRS_2	0.5	3.61	Unknown function
101053	30924	Copia_12	31	183.32	NAD(P)-binding

Table S3. Percentage of TE content in 18 fungal species

Classification	*Botrytis cinerea*		*Fusarium*		*Cryptococcus neoformans*		*Panerochaete*		*Pleurotus ostreatus*		*Serpula lacrymans*		*Puccinia*		*Pseudozyma*		*Laccaria bicolor*	*Mixia osmundae*
	B05.10	**T4**	***oxysporum***	***graminearum***	**H99**	**JEC21**	***chrysosporium***	***carnosa***	**PC15**	**PC9**	**7.3**	**7.9**	***graminis***	***striiformis***	***antarctica T34***	***hubeiensis SY62***		
DNA transposons																		
Academ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0
EnSpm	0.0	0.0	0.0	0.0	0.3	0.4	0.0	0.1	0.0	0.0	0.3	0.5	0.2	0.1	0.0	0.0	0.3	0.0
Crypton	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0
hAT	0.0	0.0	1.4	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.1	0.1	0.8	0.5	0.0	0.0	0.2	0.0
MuLE	0.0	0.0	0.8	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	2.2	0.5	0.0	0.0	0.0	0.0
PIF-Harbinger	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.4	0.5	1.3	0.5	0.0	0.0	0.3	0.0
PiggyBac	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tir1-Mariner	0.3	0.5	2.4	0.1	0.2	0.1	0.1	0.1	0.0	0.0	0.6	0.7	0.6	0.3	0.0	0.0	0.7	0.0
Helitron	0.0	0.0	0.3	0.0	0.1	0.0	0.1	0.1	0.3	0.1	0.2	0.3	0.8	0.3	0.0	0.0	0.8	0.0
Non-LTR retrotransposons																		
CR1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CRE	0.0	0.0	0.1	0.0	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jockey	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
L1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Penelope	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Tad1	0.0	0.0	1.6	0.5	0.0	0.0	0.2	0.2	0.1	0.2	0.1	0.1	0.8	0.2	0.0	0.0	1.0	0.0
LTR-retrotransposons																		
Copia	0.3	0.3	1.9	0.0	0.3	0.4	1.3	1.8	0.7	0.2	8.4	9.3	8.5	4.4	0.0	0.0	2.3	0.0
Gypsy	0.5	0.8	0.7	0.0	2.7	3.6	4.9	4.2	5.0	1.9	15.6	18.1	9.7	3.6	0.0	0.0	4.2	0.0
DIRS	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.3	0.3	0.8	0.3	0.0	0.0	0.9	0.0
BEL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Endogenous Retrovirus	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Unclassified	1.0	1.1	6.4	0.8	1.9	1.4	3.4	11.4	3.8	2.4	7.0	8.3	17.3	10.8	0.1	0.1	27.2	1.4

Low_complexity	0.6	1.0	0.2	0.2	0.4	0.4	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.3	0.0	0.2	0.2	0.0
Satellite	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Simple_repeat	0.9	1.0	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.1	0.1	0.1	0.3	0.2	0.7	0.3	0.1	0.1
Genome size (Mb)	42.7	41.6	61.4	36.5	18.9	19.1	35.2	46.3	34.3	35.6	47.0	42.7	88.6	64.8	18.1	18.4	60.7	13.6
Percentage of Genome size (known TE families)	1.2	1.6	9.4	0.6	4.0	5.3	6.8	7.0	6.1	2.5	26.0	29.8	26.0	10.7	0.0	0.0	10.7	0.0
Percentage of Genome size (all TE families)	2.2	2.7	15.8	1.4	5.9	6.6	10.2	18.4	9.9	4.9	33.0	38.1	43.3	21.6	0.1	0.1	37.9	1.5

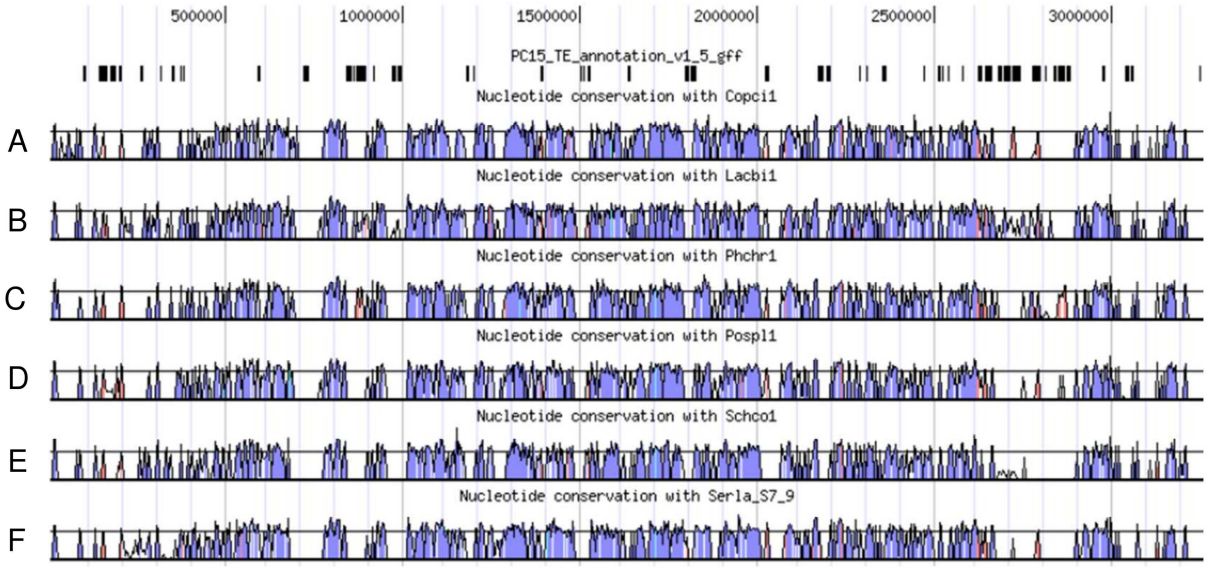

Figure S1. TE-mediated loss of conservation between *P. ostreatus* and other basidiomycetes on chromosome VII (A = *Coprinopsis cinerea*, B = *Laccaria bicolor*, C = *Phanerochaete chrysosporium*, D = *Postia placenta*, E = *Schizophyllum commune*, F = *Serpula lacrymans*). TEs are shown as black rectangles in the upper part of the panel.

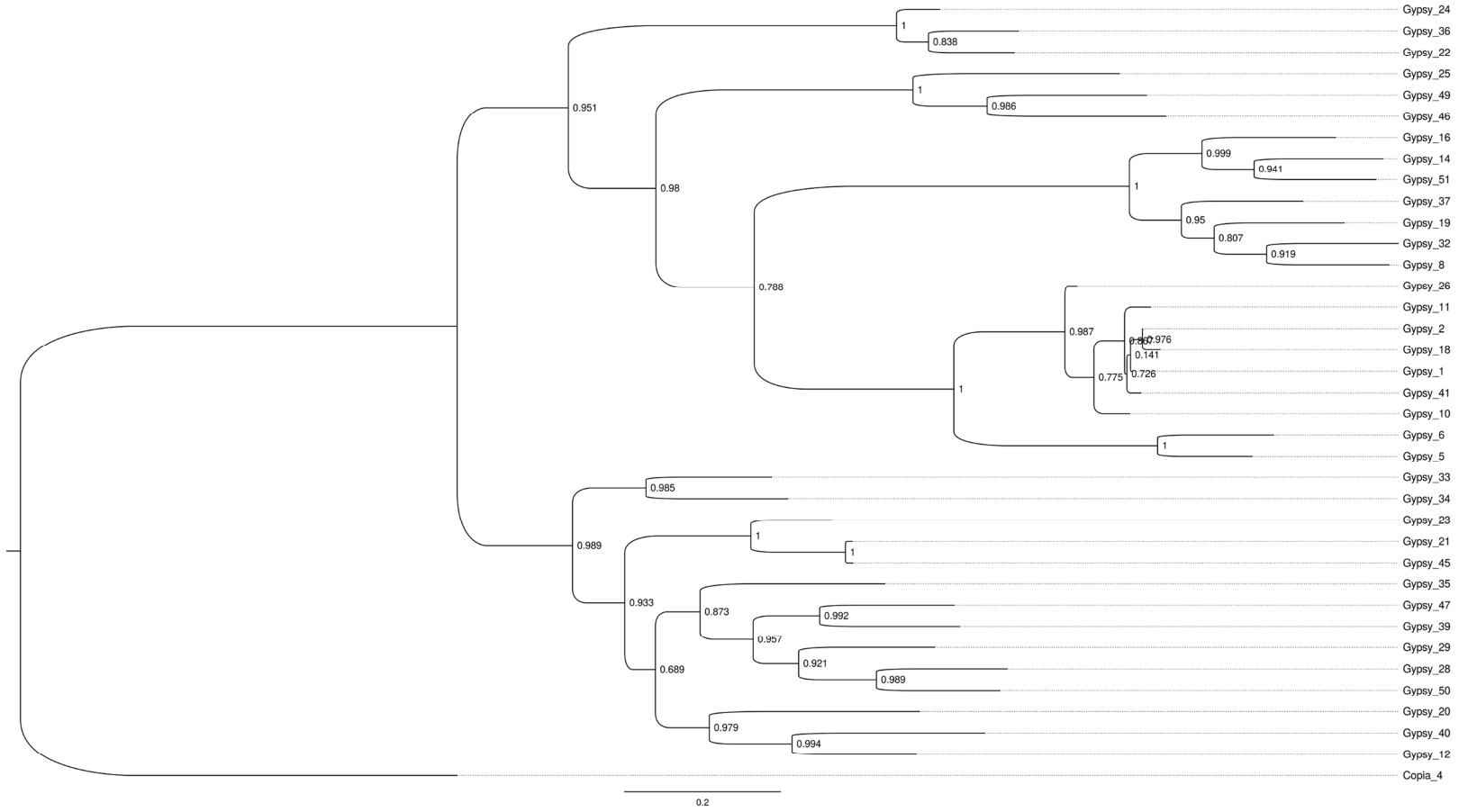

Figure S2. Phylogenetic reconstruction of Gypsy LTR-retrotransposons

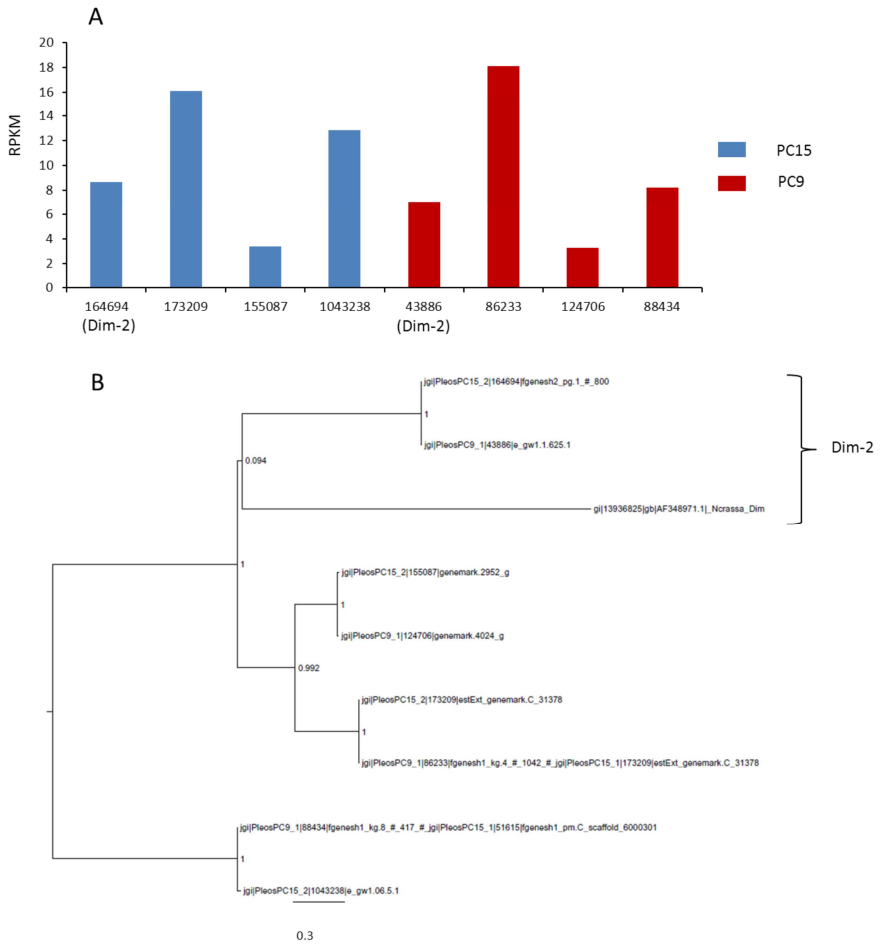

Figure S3. Expression (A) and phylogeny (B) of *P. ostreatus* DNA methyltransferases. *"Ncrassa_dim"* in panel B represents the sequence of *Neurospora crassa* Dim-2.

References

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. doi: 10.1093/bioinformatics/btp348

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi: 10.1093/nar/gkh340

Marchler-Bauer A, Bryant SH (2004) CD-Search: Protein domain annotations on the fly. Nucleic Acids Res. doi: 10.1093/nar/gkh454

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650. doi: 10.1093/molbev/msp077

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. doi: 10.1186/1471-2105-6-3

6.3. Chapter IV

Genome sequencing and annotation of the basidiomycete Coniophora olivacea

Table S1. Summary of the annotation of CAZY families in *C. olivacea* and *C. puteana*

CAZY Family	***C. olivacea***	***C. puteana***
AA	49	51
AA1 Auxiliary Activity Family 1	8	8
AA1 Auxiliary Activity Family 1	1	1
AA1_1 Laccase	6	6
AA1_2 Ferroxidase	1	1
AA3 Auxiliary Activity Family 3	25	21
AA3_1 Cellobiose dehydrogenase	2	2
AA3_2 GMC oxidoreductase	17	14
AA3_3 Alcohol oxidase	5	5
AA3_4 Pyranose oxidase	1	0
AA5 Auxiliary Activity Family 5	5	6
AA5_1 Copper radical oxidase	5	6
AA6 Auxiliary Activity Family 6	2	2
AA8 Auxiliary Activity Family 8	3	4
AA9 Auxiliary Activity Family 9	6	10
CBM	28	27
CBM1 Carbohydrate-Binding Module Family 1	2	2
CBM12 Carbohydrate-Binding Module Family 12	1	1
CBM13 Carbohydrate-Binding Module Family 13	3	6
CBM18 Carbohydrate-Binding Module Family 18	1	1
CBM20 Carbohydrate-Binding Module Family 20	2	2
CBM21 Carbohydrate-Binding Module Family 21	2	2
CBM38 Carbohydrate-Binding Module Family 38	0	1
CBM43 Carbohydrate-Binding Module Family 43	1	1
CBM48 Carbohydrate-Binding Module Family 48	3	3
CBM5 Carbohydrate-Binding Module Family 5	8	7
CBM50 Carbohydrate-Binding Module Family 50	5	1
CE	19	20
CE16 Carbohydrate Esterase Family 16	6	7
CE4 Carbohydrate Esterase Family 4	9	9
CE5 Carbohydrate Esterase Family 5	1	1
CE8 Carbohydrate Esterase Family 8	2	2
CE9 Carbohydrate Esterase Family 9	1	1
EXPN Distantly related to plant expansins	22	19
GH	215	242
GH1 Glycoside Hydrolase Family 1	3	3
GH10 Glycoside Hydrolase Family 10	3	3
GH114 Glycoside Hydrolase Family 114	0	1
GH115 Glycoside Hydrolase Family 115	2	2
GH12 Glycoside Hydrolase Family 12	5	4
GH125 Glycoside Hydrolase Family 125	1	1
GH128 Glycoside Hydrolase Family 128	7	9
GH13 Glycoside Hydrolase Family 13	6	6
GH13_1 Glycoside Hydrolase Family 13	1	1
GH13_25 Glycoside Hydrolase Family 13	1	1
GH13_32 Glycoside Hydrolase Family 13	1	1
GH13_40 Glycoside Hydrolase Family 13	1	1
GH13_5 Glycoside Hydrolase Family 13	1	1
GH13_8 Glycoside Hydrolase Family 13	1	1

GH131 Glycoside Hydrolase Family 131	2	2
GH133 Glycoside Hydrolase Family 133	1	1
GH15 Glycoside Hydrolase Family 15	2	2
GH16 Glycoside Hydrolase Family 16	26	24
GH17 Glycoside Hydrolase Family 17	3	4
GH18 Glycoside Hydrolase Family 18	21	28
GH2 Glycoside Hydrolase Family 2	5	5
GH20 Glycoside Hydrolase Family 20	4	4
GH23 Glycoside Hydrolase Family 23	1	1
GH25 Glycoside Hydrolase Family 25	1	2
GH27 Glycoside Hydrolase Family 27	3	4
GH28 Glycoside Hydrolase Family 28	15	13
GH29 Glycoside Hydrolase Family 29	4	4
GH3 Glycoside Hydrolase Family 3	13	13
GH30 Glycoside Hydrolase Family 30	5	7
GH30 Glycoside Hydrolase Family 30	0	1
GH30_3 Glycoside Hydrolase Family 30	5	6
GH31 Glycoside Hydrolase Family 31	7	12
GH32 Glycoside Hydrolase Family 32	0	1
GH35 Glycoside Hydrolase Family 35	2	2
GH37 Glycoside Hydrolase Family 37	3	4
GH38 Glycoside Hydrolase Family 38	1	1
GH43 Glycoside Hydrolase Family 43	4	6
GH45 Glycoside Hydrolase Family 45	2	1
GH47 Glycoside Hydrolase Family 47	8	9
GH5 Glycoside Hydrolase Family 5	19	21
GH5_12 Glycoside Hydrolase Family 5	3	2
GH5_15 Glycoside Hydrolase Family 5	2	2
GH5_22 Glycoside Hydrolase Family 5	2	2
GH5_30 Glycoside Hydrolase Family 5	1	1
GH5_31 Glycoside Hydrolase Family 5	1	1
GH5_5 Glycoside Hydrolase Family 5	2	5
GH5_50 Glycoside Hydrolase Family 5	1	1
GH5_7 Glycoside Hydrolase Family 5	2	3
GH5_9 Glycoside Hydrolase Family 5	4	4
GH5_dist Glycoside Hydrolase Family 5	1	0
GH51 Glycoside Hydrolase Family 51	1	3
GH53 Glycoside Hydrolase Family 53	1	1
GH55 Glycoside Hydrolase Family 55	4	5
GH6 Glycoside Hydrolase Family 6	1	2
GH63 Glycoside Hydrolase Family 63	1	1
GH7 Glycoside Hydrolase Family 7	1	2
GH71 Glycoside Hydrolase Family 71	5	6
GH72 Glycoside Hydrolase Family 72	1	1
GH76 Glycoside Hydrolase Family 76	3	3
GH78 Glycoside Hydrolase Family 78	2	2
GH79 Glycoside Hydrolase Family 79	4	4
GH81 Glycoside Hydrolase Family 81	1	1
GH85 Glycoside Hydrolase Family 85	1	1
GH88 Glycoside Hydrolase Family 88	1	1
GH89 Glycoside Hydrolase Family 89	2	2
GH9 Glycoside Hydrolase Family 9	1	1
GH92 Glycoside Hydrolase Family 92	3	4
GH93 Glycoside Hydrolase Family 93	2	1

GH95 Glycoside Hydrolase Family 95	1	1
GT	58	56
GT1 GlycosylTransferase Family 1	4	4
GT15 GlycosylTransferase Family 15	2	2
GT2 GlycosylTransferase Family 2	12	12
GT20 GlycosylTransferase Family 20	6	4
GT21 GlycosylTransferase Family 21	1	1
GT22 GlycosylTransferase Family 22	4	4
GT24 GlycosylTransferase Family 24	2	1
GT3 GlycosylTransferase Family 3	1	1
GT31 GlycosylTransferase Family 31	1	1
GT32 GlycosylTransferase Family 32	1	1
GT33 GlycosylTransferase Family 33	1	1
GT35 GlycosylTransferase Family 35	1	1
GT39 GlycosylTransferase Family 39	3	3
GT4 GlycosylTransferase Family 4	3	3
GT48 GlycosylTransferase Family 48	2	4
GT49 GlycosylTransferase Family 49	2	1
GT50 GlycosylTransferase Family 50	1	1
GT57 GlycosylTransferase Family 57	2	2
GT58 GlycosylTransferase Family 58	1	1
GT59 GlycosylTransferase Family 59	1	1
GT66 GlycosylTransferase Family 66	1	1
GT69 GlycosylTransferase Family 69	1	1
GT76 GlycosylTransferase Family 76	1	1
GT8 GlycosylTransferase Family 8	3	3
GT90 GlycosylTransferase Family 90	1	1
Myosin_motor Glycosyltransferase Family 2	3	3
PL	3	3
PL14 Polysaccharide Lyase Family 14	3	3
PL14 Polysaccharide Lyase Family 14	1	1
PL14_4 Polysaccharide Lyase Family 14	1	1
PL14_5 Polysaccharide Lyase Family 14	1	1
TOTAL	**397**	**421**

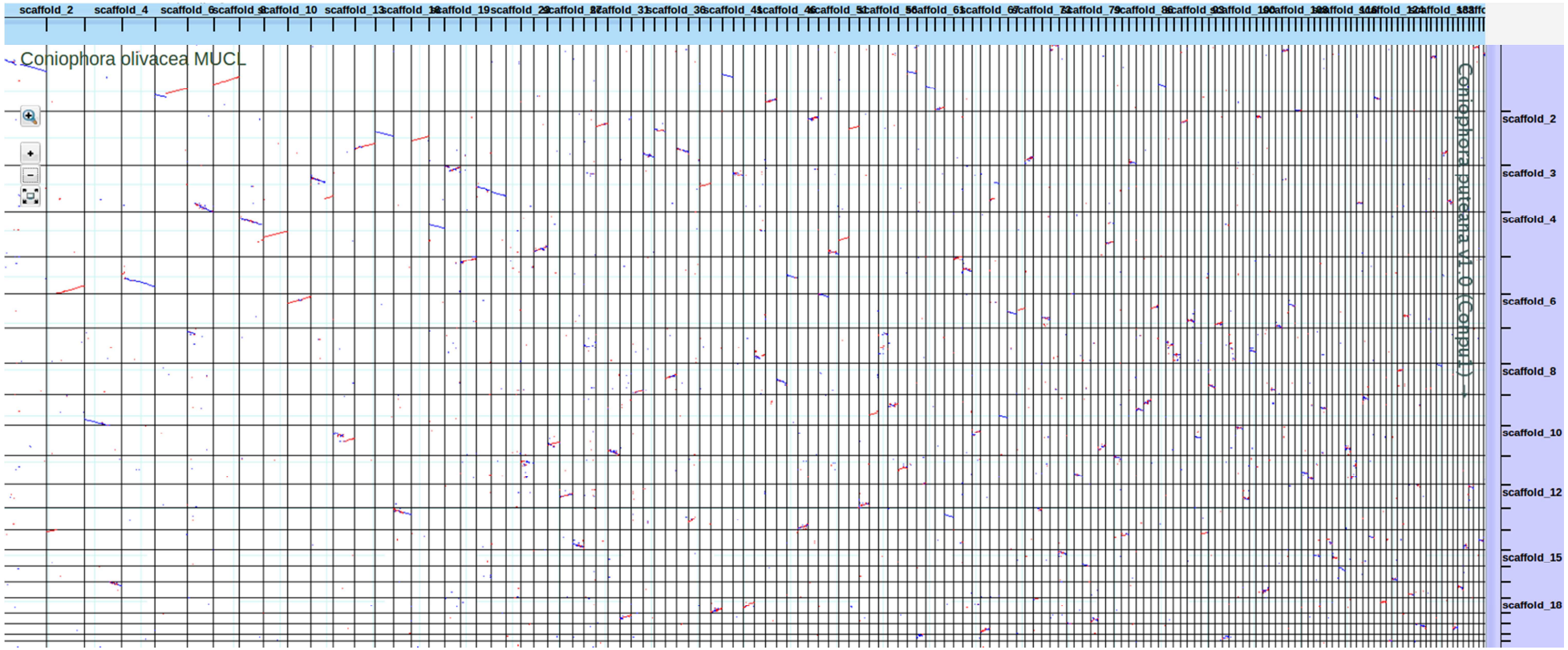

Figure S1. Snapshot of whole-genome synteny between *Coniophora olivacea* and *Coniophora puteana*

Conclusions

- *P. ostreatus* genome is populated by HELPO1 and HELPO2, two young helitron families that show differential transcriptional activity, potential ability to transpose and harbour gene-like captured sequences of uncertain origin.

- *P. ostreatus* harbours a diverse array of transposable elements that generate intraspecific diversity. The mayority of TEs are Class II elements derived from just a few families, especially LTR-retrotransposons that experienced amplification bursts during the last five million years.

- Transposable element insertions produce genome-wide repression of upstream and downstream genes in *P. ostreatus* and other basidiomycetes. This phenomenom is presumably linked to the epigenetic machinery that controls TE proliferation.

- Transposable element content is highly variable in basidiomycetes, and TE expansions are tigthly related to genome size variation. This relationship is stronger in *Pucciniomycotina* and *Ustilagionmycotina* than in *Agaricomycotina*.

- There is an important need for the fungal community to benchmark pipelines and methodologies for TE detection and build reference TE annotations, similarly to those available for protein-coding genes.

List of publications

Publications derived from this thesis:

- **Castanera R**, Pérez G, López L, Sancho R, Santoyo F, Alfaro M, Gabaldón T, Pisabarro AG, Oguiza JA, Ramírez L. Highly expressed captured genes and cross-kingdom domains present in Helitrons create novel diversity in *Pleurotus ostreatu*s and other fungi. 2014. BMC Genomics; 15(1):1071. doi: 10.1186/1471-2164-15-1071. Impact factor: 3.98
- **Castanera R**, López-Varas L, Borgognone A, LaButti K, Lapidus A, Schmutz J, et al. Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. 2016. PLoS Genet. 12:e1006108. doi: 10.1371/journal.pgen.1006108. Impact factor: 6.66
- **Castanera R,** Pérez G, López-Varas L, Haridas S, Amselem J, Grigoriev I V, Pisabarro AG, Ramírez L (2017) Comparative genomics of *Coniophora olivacea* reveals different waves of genome expansion in Boletales. Submitted
- **Castanera R**, Borgognone A, Pisabarro AG, Ramírez L. Biology, dynamics, and applications of transposable elements in basidiomycete fungi. 2017. Appl Microbiol Biotechnol. 101: 1337. doi:10.1007/s00253-017-8097-8. Impact factor: 3.37

Other contributions published during the PhD period:

- Borgognone A, **Castanera R**, Muguerza E, Pisabarro AG, Ramírez L. Somatic transposition and meiotically driven elimination of an active helitron family in *Pleurotus ostreatus*. 2017. DNA Res. doi: 10.1093/dnares/dsw060. Impact factor: 5.26
- Alfaro M, **Castanera R**, Lavín JL, Grigoriev I V., Oguiza JA, Ramírez L, et al. Comparative and transcriptional analysis of the predicted secretome in the lignocellulose-degrading basidiomycete fungus *Pleurotus ostreatus*. 2016. Environ. Microbiol. doi: 10.1111/1462-2920.13360. Impact factor: 5.93
- **Castanera R**, López L, Pisabarro AG, Ramírez L. Validation of reference genes for transcriptional analyses in *Pleurotus ostreatus* using RT-qPCR. 2015. Applied and Environmental Microbiology. doi:10.1128/AEM.00402-15. Impact factor: 3.82
- Fernández-Fueyo E, **Castanera R**, Ruiz-Dueñas F.J, López Lucendo M.F, Ramírez L, Pisabarro A.G, Martínez A.T. Ligninolytic peroxidase gene expression by *Pleurotus ostreatus*: Differential regulation in lignocellulose medium and effect of temperature and pH. 2014. Fungal Genetics and Biology. doi:10.1016/j.fgb.2014.02.003. Impact factor: 2.59
- **Castanera R**, Omarini A, Santoyo F, Pisabarro A.G, Ramirez L. Non-Additive Expression Underlies Dikaryotic Superiority in *Pleurotus ostreatus* Laccase Activity. 2013. PLOS ONE. 8(9): e73282. doi:10.1371/journal.pone.0073282. Impact factor: 3.53
- Foulongne-Oriol M, Murat C, **Castanera R**, Ramírez L, Sonnenberg A. Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus*. 2013. Fungal Genetics and Biology. doi: 10.1016/j.fgb.2013.04.003. Impact factor: 3.26
- Parenti A, Muguerza E, Redin I, Omarini A, Conde E, Alfaro M, **Castanera R**. Santoyo F. Ramirez L., Pisabarro A.G. Induction of laccase activity in the white rot fungus *Pleurotus ostreatus* using water polluted with wheat straw extracts. 2013. Bioresource Technology. doi:10.1016/j.biortech.2013.01.072. Impact factor: 5.04

Funding

Raúl Castanera obtained a FPI-PhD scholarship from the Ministry of Economy and Competitiveness (Grant number BES-2012-053928).

This work has been supported by Spanish National Research Plan and U.S. Department of Energy:

- Effect of helitrons in genome structure and transcriptional profile of *Pleurotus ostreatus*. AGL2011-30495. Funded by the Spanish Ministry of Economy, Department of Research, Development and Innovation.

- Study of the interactions between transcriptome and methylome to explain differences in growth rate and mushroom yield in the edible fungus *Pleurotus ostreatus*. AGL2014-55971-R. Funded by the Spanish Ministry of Economy, Department of Research, Development and Innovation

- Joint Genome Institute and Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (http://science.energy.gov/bso/contract-management/)